TECHNICAL PAPER

NCVER

Exploratory analysis of VET market segments

Bryan Palmer

INFORMING + INFLUENCING
THE AUSTRALIAN VET SECTOR

Publisher's note

The views and opinions expressed in this document are those of NCVER and do not necessarily reflect the views of the Australian Government, or state and territory governments. Any interpretation of data is the responsibility of the author/project team.

To find other material of interest, search VOCEDplus (the UNESCO/NCVER international database <http://www.voced.edu.au>) using the following keywords: data analysis; enrolment; international students; language; migrants; participation; social inclusion; statistical analysis; statistical method; students; vocational education and training.

This document should be attributed as Palmer, B 2022, *Exploratory analysis of VET market segments,* NCVER, Adelaide.

This work has been produced by NCVER on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Government Department of Education, Skills and Employment.

COVER IMAGE: GETTY IMAGES

ISBN 978-1-925717-96-9
TD/TNC 147.05

Published by NCVER, ABN 87 007 967 311

Level 5, 60 Light Square, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

Phone +61 8 8230 8400 **Email** ncver@ncver.edu.au
Web <https://www.ncver.edu.au> <https://www.lsay.edu.au>
Follow us: <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>

About the research

Exploratory analysis of VET market segments

Bryan Palmer

This paper summarises the exploratory quantitative analysis undertaken to investigate how vocational education and training (VET) students cluster and segment in the Australian VET market. This analysis is outlined in three sections:

- The first section focuses on 'clustering' as a technique for grouping data and the three clustering algorithms used. These are then discussed in more detail to provide some insights into how they operate. Their specific data requirements, along with their strengths and weaknesses, are also considered.
- In the next section the outputs of the clustering approaches are considered. The resultant clusters are examined to better understand them, and meaningfully label and group them into segments.
- With the insights gained from the clustering process, the final section of this paper returns to the raw data. This step was necessary to further explore (in this case, only some of) the identified market segments. Here three key market segments are explored: students in targeted English programs; students in social inclusion programs; and migrant students.

Key messages

- Two of the three clustering algorithms (k-means and agglomerative) were applied to the total VET activity (TVA) data.
- After considering the output across these two clustering algorithms, several segments within the Australian VET market were identified:
  - targeted English programs/students
  - overseas students (studying in Australia)
  - younger students (includes VET in Schools programs)
  - migrants
  - social inclusion programs/students
  - jurisdictional priorities
  - program enrolments not elsewhere identified (NEI)
  - subject only enrolments NEI.
- The VET system collects largely categorical variables — with different levels of consistency and completeness — for millions of students, programs and subjects. As a result, it is not well suited to the application of clustering algorithms. Despite this, two clustering algorithms (k-means and agglomerative) were applied to the data, with a third (DBSCAN) unable to be applied successfully.
- While clustering algorithms can carve a dataset into clusters, identifying something that is meaningful to practitioners in a way that explains the clusters is not always guaranteed. Sometimes it can be challenging to bring a useful human perspective or narrative to the clustered outputs. The approach taken in this paper was to look at the features in each cluster that were overly represented compared with all students.

- The algorithms applied assumed single cluster membership to the exclusion of all others. This is an analytically useful (but unrealistic) simplification. In real life, the identified market segments are not mutually exclusive, and students may belong to more than one segment.
- The research approach was unable to conclusively use the clustering outputs to determine whether the identified clusters align with, or bring insights to, the other typologies for segmenting the Australian VET market that can be found in academic literature.

Simon Walker
Managing Director, NCVER

Acknowledgments

The author would like to thank Adrian Ong, Patrick Lim, Charlotte Scobie, Michelle Hall, Zhenyuan Li and Melinda Lees from the National Centre for Vocational Education Research (NCVER) for their advice, suggestions, support and encouragement throughout this project. The author has benefited from his many conversations with the team at the NCVER; nonetheless, any remaining errors in the text of this report are the author's alone.

Contents

Tables and figures

Tables

Figures

Introduction

The Australian vocational education and training (VET) sector is large and complex, with some students enrolled in one or more programs, others in subjects not part of a nationally recognised program, and many in a combination of both.

In 2020, there were 3.9 million students enrolled in nationally recognised VET, and an estimated 21.7% of the Australian resident population aged 15 to 64 years participated in nationally recognised VET (NCVER 2021). A little over half of all enrolments (2.4 million) were in subjects that were not part of a nationally recognised program of study. Completing these subjects often fulfilled regulatory or other safety requirements necessary for employment, holding a particular licence, or doing a particular job in the Australian context (Palmer 2021).

Around a half of the students (2 million) were enrolled in nationally recognised qualifications. The level of education in these qualifications ranged from the pre-vocational level (certificate I) to university postgraduate level (graduate diploma). While apprenticeships are the popular public face of the VET system, only some 300 000 students (around 7.6% of all students) were enrolled in a program of study that was also part of an apprenticeship or traineeship.

Adding to this complexity, the VET system is also deployed in a range of niche contexts. For example, for students disengaged from the school system, it provides many with a second chance at education; it is used to teach English to migrants; and it is used by governments to help re-engage people with the world of work, especially younger people who are otherwise disconnected from the labour market.

Individuals not only engage with the VET system immediately after school to prepare for a career, but many also return or engage for the first time in mid-life or mid-career — or even later in life — to seek out new career paths and new opportunities.

Given this diversity of products and purposes, how should the complexity of the VET market be comprehended and mapped? This is the question the author sought to answer, although this question could also be framed as: How do VET students in Australia cluster? Or: What are the market segments in the overall Australian VET market?

Additionally, this work set out to compare the segments identified in this research with other systems of categorisation in the VET sector, such as the learner-centred/type-of-learning matrix proposed by Circelli and Stanwick (2020), and the various roles of VET proposed by Moodie et al. (2015). However, answering these questions proved challenging.

This exploratory analysis uses total VET students and courses data from 2019 (NCVER 2020), also known as total VET activity (TVA). The 2019 dataset was used to avoid any artefacts that might arise in the 2020 TVA dataset resulting from the behavioural responses of students and the policy responses of governments to the global COVID-19 pandemic.[1]

[1] The total number of students in 2020 was down 6.4% on 2019 (NCVER 2021).

In addition to exploring the data through cluster analysis, the utility of various methodological approaches was tested in the context of VET, with the question: How useful are clustering algorithms from the discipline of unsupervised machine learning in identifying market segments in the VET student market?

Three clustering algorithms were explored: k-means, agglomerative clustering, and density-based spatial clustering of applications with noise (DBSCAN). Each algorithm has different assumptions about what a cluster looks like and how they are detected.

Clustering approaches

This section focuses on ‘clustering’ as a technique for grouping data and describes the three clustering algorithms used and some of the challenges they presented.

Supervised versus unsupervised learning: classification versus clustering

Machine learning as a discipline makes a distinction between supervised learning and unsupervised learning. Supervised learning algorithms are initially trained with datasets where the desired outcome is known and labelled within a training dataset. The algorithms learn from this training data before they are applied to data where the outcome is not known.

Classification is one type of supervised learning. For example, supervised machine learning algorithms can be trained to identify email spam, ensuring that the email program places the spam in a folder separate from the inbox (Burkov 2019, p.19).

Unsupervised learning does not have a training component; rather, the algorithm undertakes its task without the benefit of predetermined labels or a labelled training dataset.

Although clustering is analogous with classification, it is an unsupervised approach to machine learning. The final categories are not provided to the algorithm. The algorithm decides, based on the structure of the available data and the rules of the algorithm, which cases in the data are ‘like’ each other. There are many different clustering algorithms, each with their own innate view about what constitutes a cluster and how they are detected in a dataset (Geron 2019).

Unsupervised clustering can be very useful in exploratory analysis. It is used to identify structures and cleavages in the data that would be otherwise unseen, and it is often used commercially to segment customers. Moreover, it is frequently used to prompt further research questions once clusters have been identified within the data.

That said, it can also be frustrating. It is not always obvious why an algorithm has (or has not) grouped cases into a cluster. In the absence of an ideal system of categorisation (a grounded truth), it can be difficult to assess the performance of the various approaches to clustering or the outputs of a clustering algorithm. Likewise, it can be difficult to determine whether the clusters occurred because of the factors found to be analytically interesting or for other reasons (Burkov 2019; Müller & Guido 2017).

Selected clustering algorithms

According to the Statistics and Machine Learning with R GitHub repository, there are more than 100 clustering algorithms to choose from, although, arguably, only a few broad categories or types of clustering:

- centroid/partitioning approaches (for example, the k-means algorithm)
- distributional approaches (for example, Gaussian mixture model [GMM])
- connectivity/hierarchical (for example, divisive/agglomerative clustering)
- density approaches (for example, DBSCAN/OPTICS/mean-shift).

In this paper, three specific clustering algorithms were examined: k-means, agglomerative hierarchical clustering and DBSCAN. The clustering approaches were implemented using the scikit-learn libraries for the Python 3.9 programming language (Pedregosa et al. 2001).

An overview of each clustering algorithm follows, with the aim of providing some insights into their operation and to identify some of the challenges in the use of these algorithms.

K-means

The k-means algorithm requires the data analyst to specify how many clusters (k) it should find. The data are provided to the algorithm as a matrix of real numbers, where the rows are cases (students), and the columns are variables. The algorithm begins by 'guessing' k centres or centroids for the data. Each data point is then clustered based on its closest centroid. Once the data points have been clustered, the actual centroid for each cluster of data points is calculated. The algorithm is repeated with the updated centroids until no further movement occurs in the centroids between one iteration and the next.

K-means is one of the simplest and most frequently used clustering algorithm. Although it is typically a fast algorithm, it does not always find the optimal result. It needs to be run multiple times with different starting-point guesses, although, fortunately, the scikit-learn library takes care of this. In addition, the k-means algorithm often does not work well when the clusters are of different sizes, have different densities, or have non-spherical shapes. In practice, the clusters found by k-means algorithms are frequently of a similar size (Geron 2019; Müller & Guido 2017; Marsland 2015).

The issue of cluster size is significant, as we do not expect the clusters of VET students to necessarily be of similar sizes. Our solution to this problem is to search for a larger number of clusters (a higher value of k), recognising that clusters may need to be amalgamated after the clustering algorithm has been applied.

Because the k-means algorithm uses the Euclidean distance between data points, it is sensitive to the scale of the input data (Raschka & Mirjalili 2019). To address this sensitivity, every input variable must be scaled to the same scale, a practice known as feature scaling. The approach to feature scaling used in this report is discussed in appendix A.

Agglomerative hierarchical clustering

As with k-means, the agglomerative algorithm requires the analyst to specify how many clusters (k) it should find. The data can be provided to the algorithm in two forms, either as:

- a matrix of real numbers, where the rows are cases (students) and the columns are variables, or
- a pre-computed NxN distance matrix between each of the cases (students) in the dataset.

The agglomerative algorithm begins by assuming each data point is its own cluster. It then finds the two clusters that are closest to each other. These two closest clusters are merged to form a new cluster. This step reduces the total number of clusters by one. The algorithm is repeated until there are only k clusters remaining (Raschka & Mirjalili 2019).

The biggest challenge with agglomerative clustering is the space and time complexity of the algorithm. Whereas the memory space and computational time grow linearly with the number of cases being analysed using K-means, both are quadratic (of order N2) with agglomerative clustering. With millions of students and training activity records, the internal distance was too large for Python to retain in memory. To address this constraint, we limited the agglomerative hierarchical clustering to a random sample of 60 000 VET students from the 2019 TVA dataset.

As with the k-means algorithm, because Euclidean distances are used, if a distance matrix is not provided, then feature scaling is an important requirement for this algorithm.

DBSCAN

The key idea behind DBSCAN is that clusters are densely populated with data points, with the area between clusters less populated. Rather than specify the number of desired clusters, the analyst specifies two parameters: epsilon (a maximum distance for a data point to be associated with another data point) and minimum samples (the minimum number of data points required to be within the epsilon radius for a point to be defined as a ‘core point’). As with agglomerative clustering, the data for DBSCAN can be provided in one of two forms:

- a matrix of real numbers, where the rows are cases (students) and the columns are variables, or
- a pre-computed NxN distance matrix between each of the cases (students) in the dataset.

The DBSCAN algorithm begins by identifying core points: that is every data point which has at least ‘minimum samples’ data points within the (n-dimensional hyper-sphere) radius of epsilon units. A border point has less than ‘minimum samples’ data points within the radius of epsilon units, but at least one of those points is a core point. Data points that are neither core points nor border points are left as ‘noise’ and not clustered. Clusters are defined by the set of core points that can be traversed by a path comprising only core points, plus any connected border points.

A key advantage of DBSCAN is that clusters can take on arbitrary, non-spherical shapes. Another advantage is that it does not require the analyst to specify the number of clusters (k) beforehand. (Raschka & Mirjalili 2019; Müller & Guido 2017). However, choosing good values for the epsilon and minimum samples hyper parameters can be challenging (Burkov 2019).

Like the agglomerative algorithm, the space complexity of the DBSCAN algorithm is N2. Consequently, we limited the input to the DBSCAN algorithm to the same randomly selected cohort of 60 000 VET students.

Furthermore, as with the two previous algorithms, if a precomputed distance matrix is not provided, feature scaling is important for the same reasons.

Clustering results, and the identification of market segments

Data preparation

The 2019 TVA dataset required substantial preparation before it could be used by the clustering algorithms. The data-preparation process is discussed in some detail in appendix A.

K-means and agglomerative clustering

For the two algorithms where the value of k (the number of clusters) needs to be specified a priori, 8 and 16 clusters were opted for. These numbers were selected to ensure a smallish number of segments. We were also conscious that the k-means algorithm tends to return clusters that are similarly sized. However, we did not expect the actual clusters in the data to be similarly sized, anticipating that we may need to group clusters together when interpreting the results from the k-means algorithm.

The two clustering algorithms produce a similar output. For every one of the 5.5 million students, the k-means algorithm allocates a number (from zero to k-1) to indicate their cluster membership. Similarly, for each member of the 60 000 random sample, the agglomerative clustering algorithm allocates a number from zero to k-1. The most challenging step in cluster analysis is interpreting these results.

We begin by comparing the sizes of the clusters from the two algorithms. In table 1, showing percentage sizes (of the union set across the two approaches), at k = 8 we can see that the k-means algorithm (in the row totals) produced broadly similar-sized clusters, whereas the agglomerative algorithm produced markedly different-sized clusters. We can also see there is some overlap, but also differences between the two clustering algorithms when the clusters are cross-tabulated. It is a similar story with k = 16.

Table 1 Comparison of cluster sizes for agglomerative and k-means algorithms (%)

Agglomerative, k = 8	0	1	2	3	4	5	6	7	Total
K-means, k = 8									
0	0.0	21.6	0.4	0.0	0.2	0.1	0.0	1.0	**23.3**
1	0.1	0.4	2.5	0.1	5.5	0.8	2.2	0.0	**11.7**
2	0.0	13.9	0.3	0.0	0.0	0.1	0.0	1.0	**15.3**
3	4.3	0.0	0.0	0.0	0.5	0.0	0.0	0.0	**4.8**
4	0.1	0.9	0.2	0.4	9.8	0.1	1.2	0.0	**12.7**
5	0.0	13.7	0.7	0.0	0.1	0.1	0.0	1.1	**15.7**
6	0.1	0.0	0.0	6.1	0.0	0.0	0.2	0.0	**6.4**
7	0.1	1.5	0.1	0.3	6.4	0.0	1.7	0.0	**10.1**
Total	**4.8**	**52.1**	**4.1**	**7.0**	**22.7**	**1.1**	**5.3**	**3.1**	**100.0**

To interpret the clusters, we will highlight any feature where:

- on a threshold basis, if more than 90% of the feature was of a particular value
- on a multiplicative basis, if the proportion for a cluster is greater than three times the mean for the total population of students

- on an additive basis, if the proportion for a cluster is more than 15 percentage points over the mean for the total population of students.

The reason for having both additive and multiplicative approaches for feature selection is to account for different-sized populations. For example, looking at the state in which the training was delivered: if New South Wales is the delivery site for roughly 35% of all VET students, it will be identified as a feature for a cluster when more than 50% of the students in the cluster come from New South Wales. If the Northern Territory is the delivery site for roughly 1% of all students, it will be identified as a feature when 3% or more of students in a cluster come from the Northern Territory. A cluster may have more than one feature for the same data item. In our example, it is theoretically possible for both New South Wales and the Northern Territory to be identified as a feature.

To consolidate these clusters, each with a long list of features, into market segments, the following heuristic was applied:

- If over 50% of students in the cluster were undertaking a targeted English program of study, then the cluster was labelled as being a 'Targeted English' segment.
- If the cluster does not have a segment label, and it has the Remoteness = Overseas feature, then the cluster was labelled as an 'Overseas' student segment.
- If the cluster does not have a segment label, and it has the Younger = True feature, then the cluster was labelled as 'Younger' students.
- If the cluster does not have a segment label, and it has the Born Overseas feature, then it was labelled as 'Migrants' segment.
- If the cluster does not have a segment label, and it has the Unemployed, or Disability or Indigenous or targeted English features, it was labelled as 'Social inclusion' segment.
- If the cluster does not have a segment label, and a delivery location was featured, then the cluster was labelled as 'Jurisdictional priorities'. Jurisdictional priorities can emerge both in the private sector, because of the different demands for skills in each of the state economies, and in the public sector, as a result of the different policy objectives of, and funding incentives from, state governments.
- If the cluster does not have a segment label, and more than 80% of the segment was enrolled in a program (be it a nationally recognised qualification, an accredited qualification, an accredited course, a locally developed course, or a locally developed skill set), it was placed in the 'Program enrolments not elsewhere identified (NEI)' segment.
- If the cluster does not have a segment label, and more than 80% of the segment is subject only enrolments, the cluster was labelled as a 'Subject only enrolments NEI' segment.
- Finally, if the cluster does not have a segment label, it was labelled as 'Mixed enrolments NEI' segment. As it happened, this catch-all category was not needed.

The idea behind this stepwise heuristic was to assign a cluster to a smaller, more specifically designated, segment first, if that was at all possible. And only if it could not be segmented at a detailed level, it would be grouped more generally. These segments were identified after considering the list of features from the output of the clustering process. We acknowledge that this is not the only possible arrangement of market segments to emerge from the clustering of the 2019 TVA data and that other approaches to identifying market segments could be taken.

The detailed list of features for the individual clusters across the four clustering approaches (k-means and agglomerative, and at k = 8 and k = 16) are set out in appendix B. These tables also identify for each cluster all the possible segment names that could have been applied, and the segment name that was actually applied.

Summarising the segments

Because this is an exploratory analysis, we will be working with all four outputs rather than working with only one of the four outputs from the clustering analysis. Combining the different outputs provides a richer picture for each of the identified market segments; it also allows us to identify when different approaches produce similar clusters and segments.

Table 2 sets out the number of clusters (in cells) allocated to each segment (rows) for each of the four clustering approaches (columns).

Table 2 The number of clusters allocated to each segment by clustering approach

	K-means, k = 8	K-means, k = 16	Agglomerative, k = 16	Agglomerative, k = 8
Jurisdictional priorities	0	4	2	1
Migrants	0	1	2	1
Overseas students	1	1	1	1
Program enrolments NEI	2	4	2	1
Social inclusion	1	0	1	1
Subject only enrolments NEI	3	4	2	1
Targeted English	0	1	1	0
Younger students	1	1	5	2

The indicative size of these market segments in terms of the percentage of students in the segment is set out in table 3. Not all market segments were seen in all approaches.

Table 3 The indicative size of market segments (%)

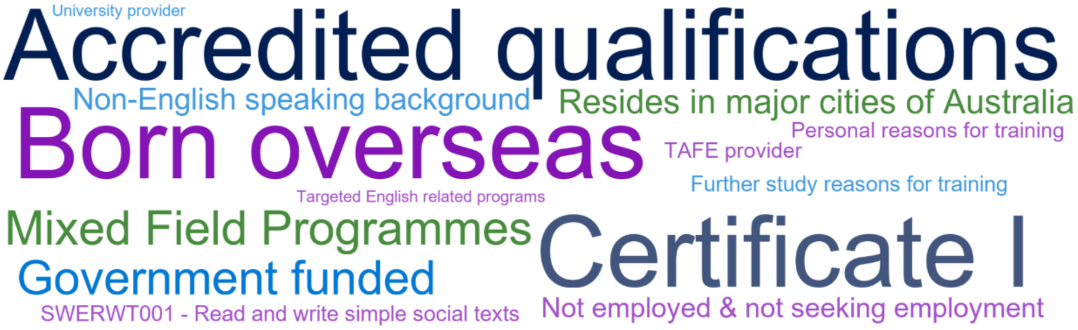
	K-means, k = 8	K-means, k = 16	Agglomerative, k = 16	Agglomerative, k = 8
Jurisdictional priorities	-	18.7	3.9	3.1
Migrants	-	3.2	3.0	4.1
Overseas students	4.9	4.5	4.5	4.8
Program enrolments NEI	22.8	31.5	19.9	22.7
Social inclusion	11.7	-	2.4	1.1
Subject only enrolments NEI	54.2	33.1	52.1	52.1
Targeted English	-	3.0	1.7	-
Younger students	6.4	6.2	12.5	12.2
% Total	**100.0**	**100.0**	**100.0**	**100.0**

To better understand each segment, we will now consider a word cloud to illustrate the segment features and a frequency histogram to show the top features for each segment.

Targeted English

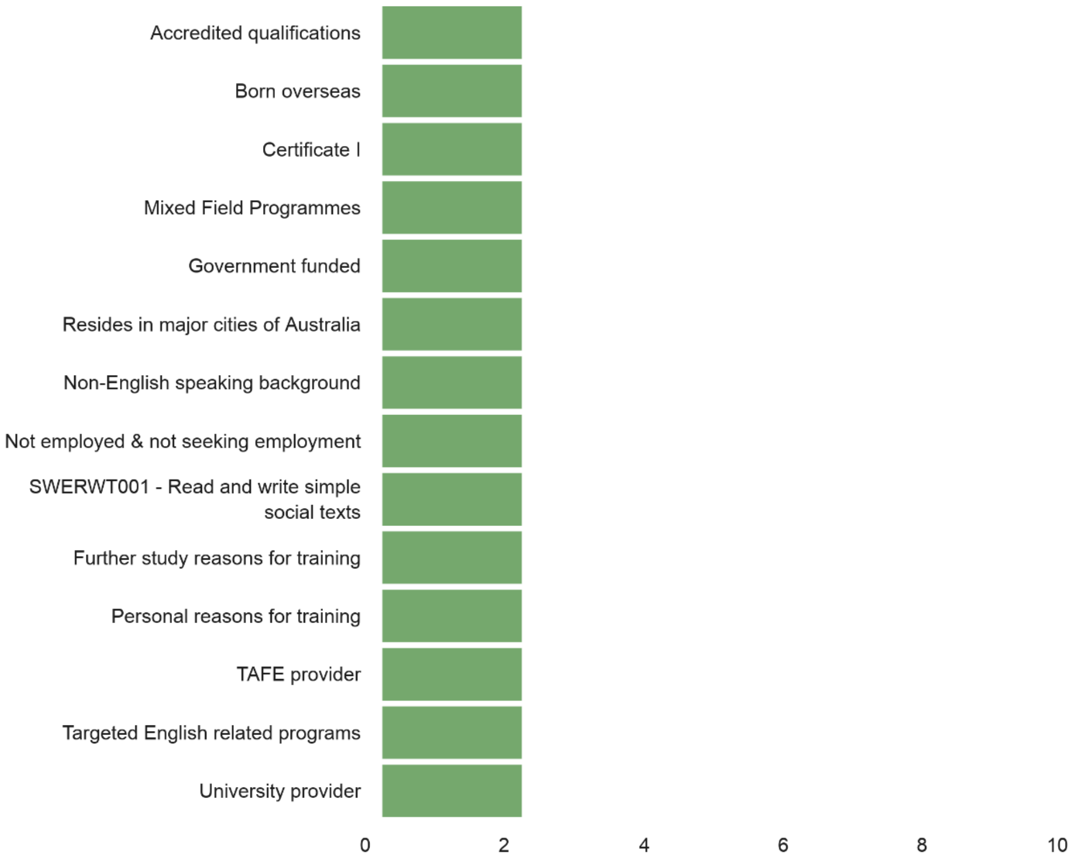
The targeted English segment only appeared in the output from the two algorithms when k was set to 16, and in both algorithms it was a small proportion of the overall VET market. The most popular subject in this cluster was SWERWT001 - Read and write simple social texts. The training was often provided as an accredited qualification. TAFE (technical and further education) institutes and universities were a key provider of the training, with many students being government-funded. Many students in this segment were born overseas.

Figure 1 Targeted English word cloud



Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 2 Targeted English top features

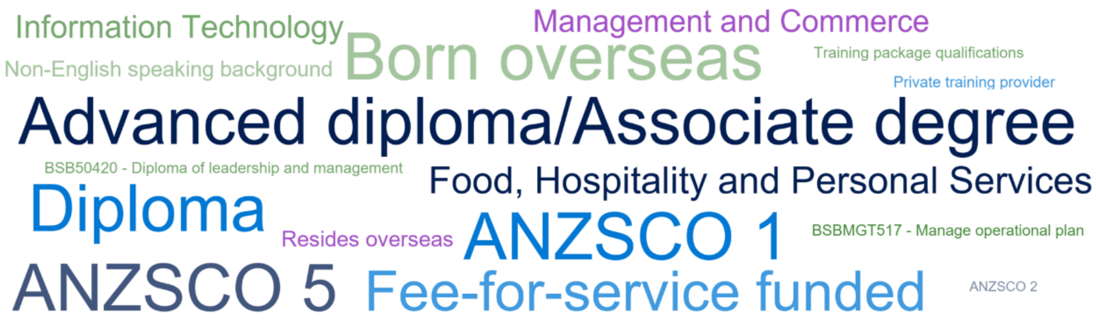


Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

Overseas students

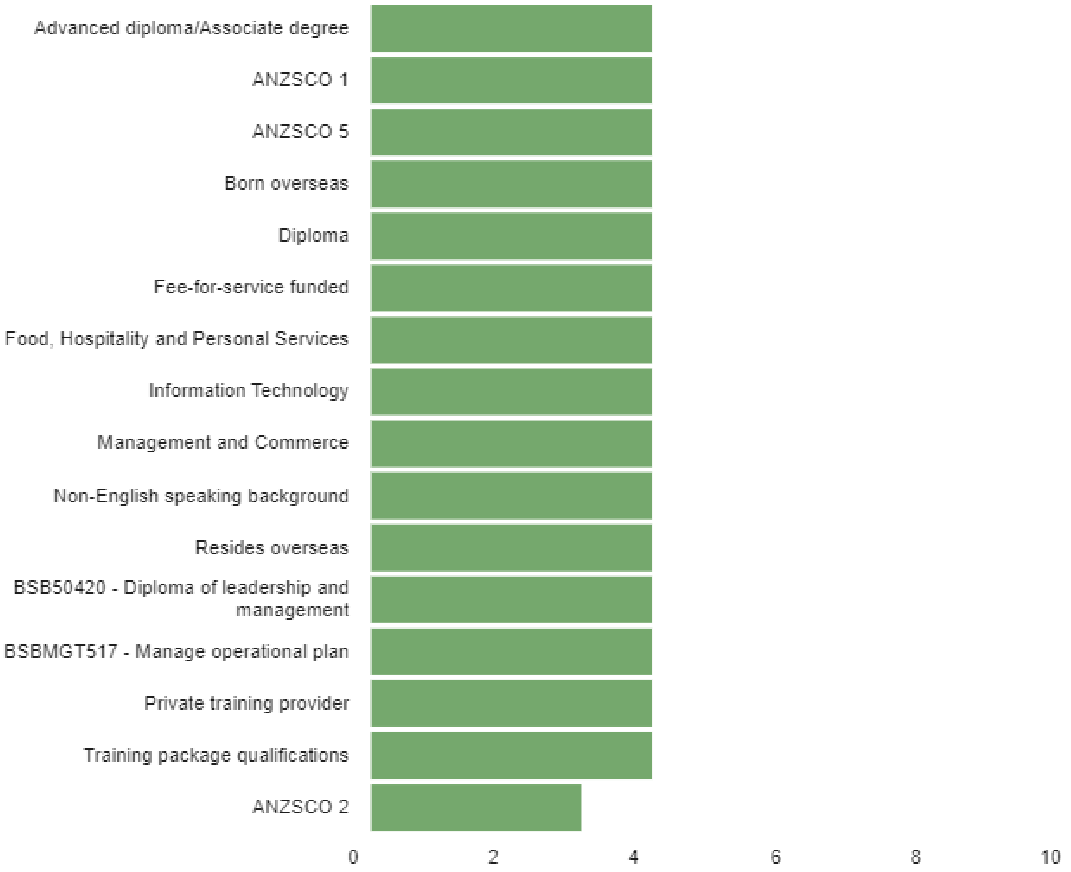
The Overseas student segment appeared in the output of the four clustering algorithms. These programs of study were typically provided on a fee-for-service basis, and the level of education was often at the diploma and advanced diploma levels, with the providers often private training providers. The most popular program was BSB50420 - Diploma of Leadership and Management. The most popular unit of competency was BSBMGT517 - Manage operational plan. In addition to Management and Commerce, Information Technology, and Food, Hospitality and Personal Services were featured fields of education.

Figure 3 Overseas word cloud[2]


Information Technology
Management and Commerce
Non-English speaking background
Born overseas
Training package qualifications
Private training provider
Advanced diploma/Associate degree
BSB50420 - Diploma of leadership and management
Food, Hospitality and Personal Services
Diploma
Resides overseas
ANZSCO 1
BSBMGT517 - Manage operational plan
ANZSCO 5
Fee-for-service funded
ANZSCO 2

Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 4 Overseas top features[2]



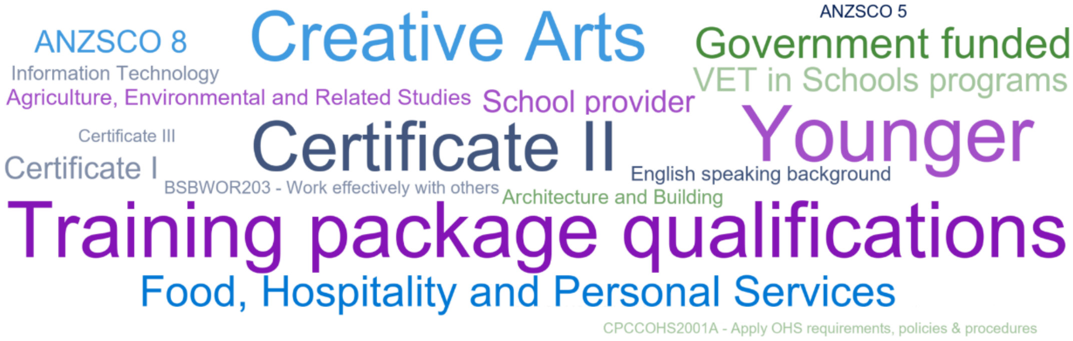
Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

[2] The ANZSCO classification is a code that uniquely identifies the type of occupation which a client would be qualified in on completion of their training. ANZSCO 1 refers to Managers, ANZSCO 2 to Professionals and ANZSCO 5 to Clerical and Administrative Workers.

Younger students

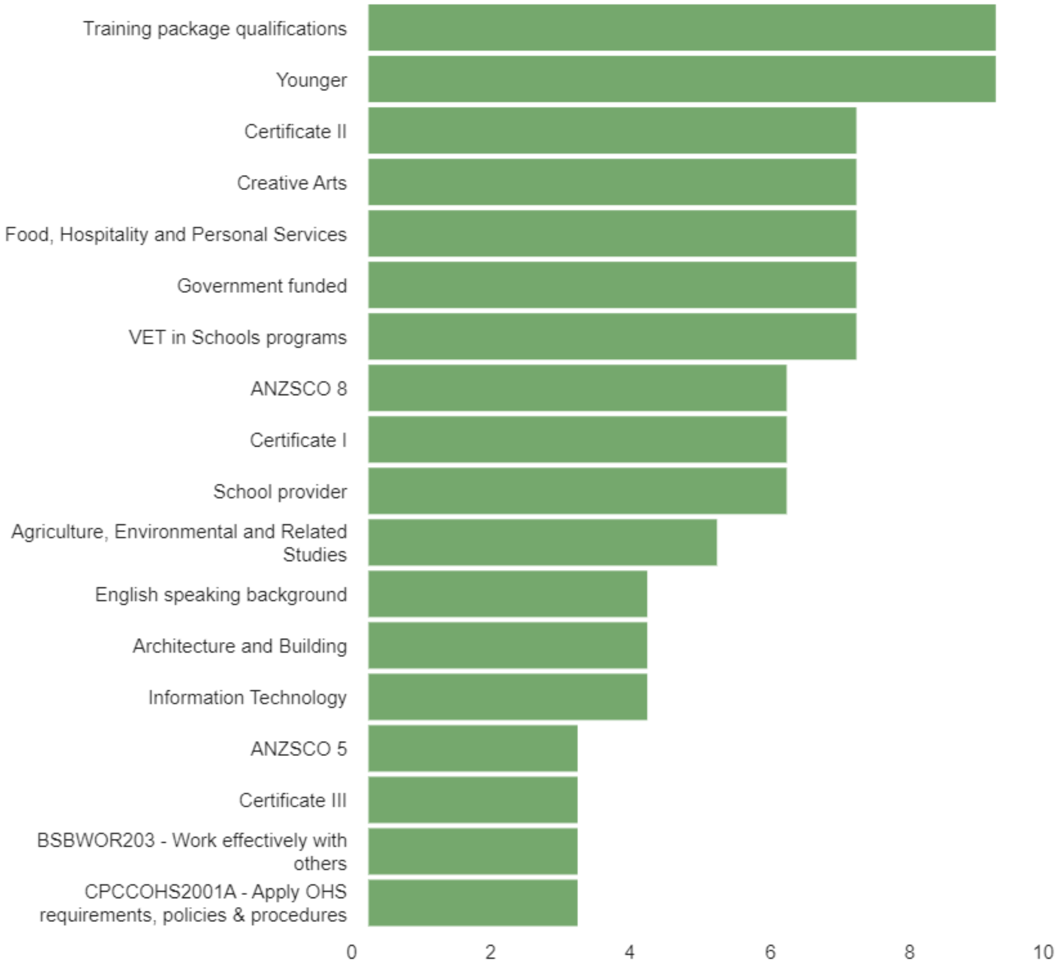
Programs for younger students were found by all four algorithms. These were associated with VET in Schools programs and delivery at schools. The level of education was typically in the certificate I to III range, and the training was typically government-funded.

Figure 5 Younger students word cloud[3]



Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 6 Younger students top features[3]



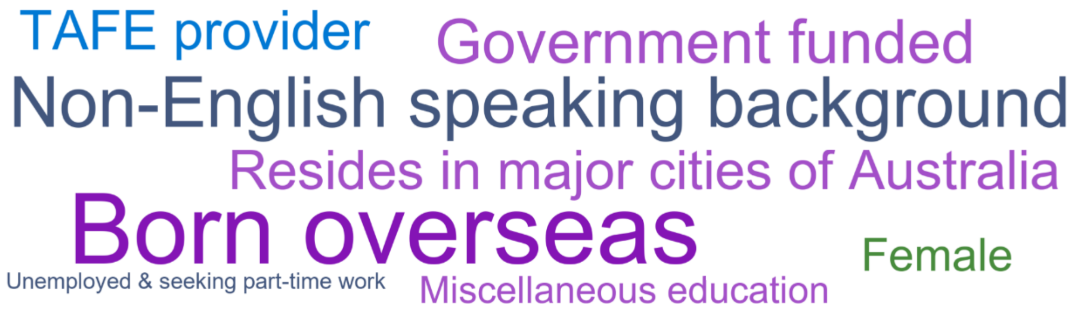
Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

[3] The ANZSCO classification is a code that uniquely identifies the type of occupation which a client would be qualified in on completion of their training. ANZSCO 5 refers to Clerical and Administrative Workers and ANZSCO 8 to Labourers.

Migrants

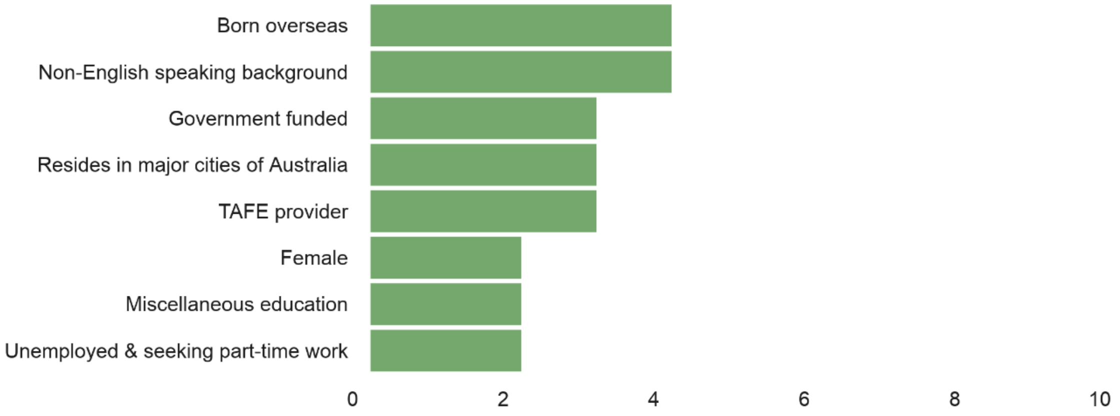
The migrants segment was found by three of the four clustering approaches. A high proportion of students in this segment were residing in major cities and delivery was often government-funded with a high proportion provided by a TAFE institute.

Figure 7 Migrants word cloud


TAFE provider
Government funded
Non-English speaking background
Resides in major cities of Australia
Born overseas
Female
Unemployed & seeking part-time work
Miscellaneous education

Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 8 Migrants top features

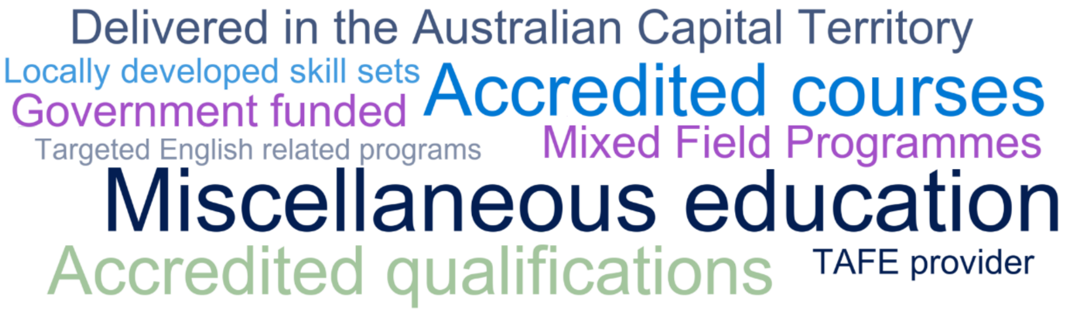


Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

Social inclusion

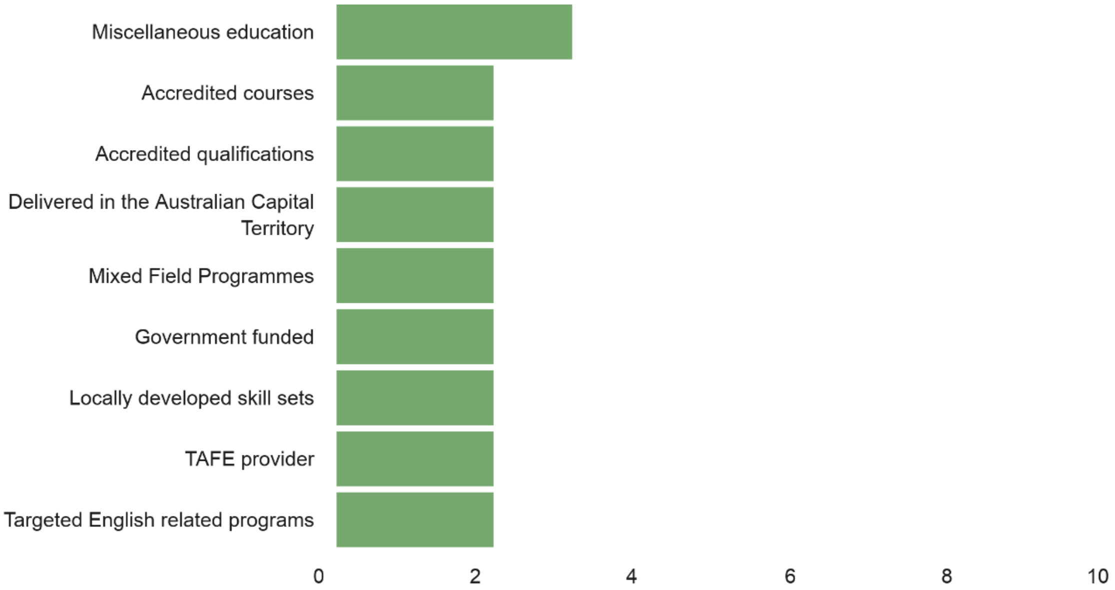
Programs designed to promote social inclusion appeared in the output from three of the clustering algorithms. The programs were often a mix of locally developed skill sets, accredited courses and accredited qualifications.

Figure 9 Social inclusion word cloud


Delivered in the Australian Capital Territory
Locally developed skill sets
Accredited courses
Government funded
Targeted English related programs
Mixed Field Programmes
Miscellaneous education
Accredited qualifications
TAFE provider

Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 10 Social inclusion top features

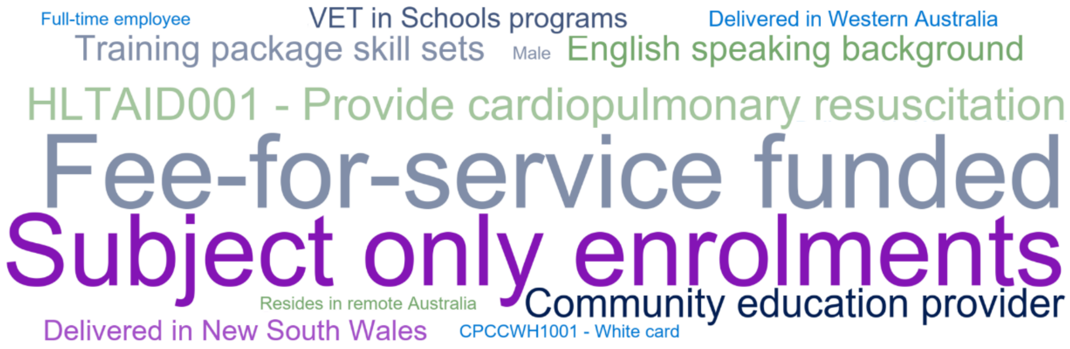


Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

Jurisdictional priorities

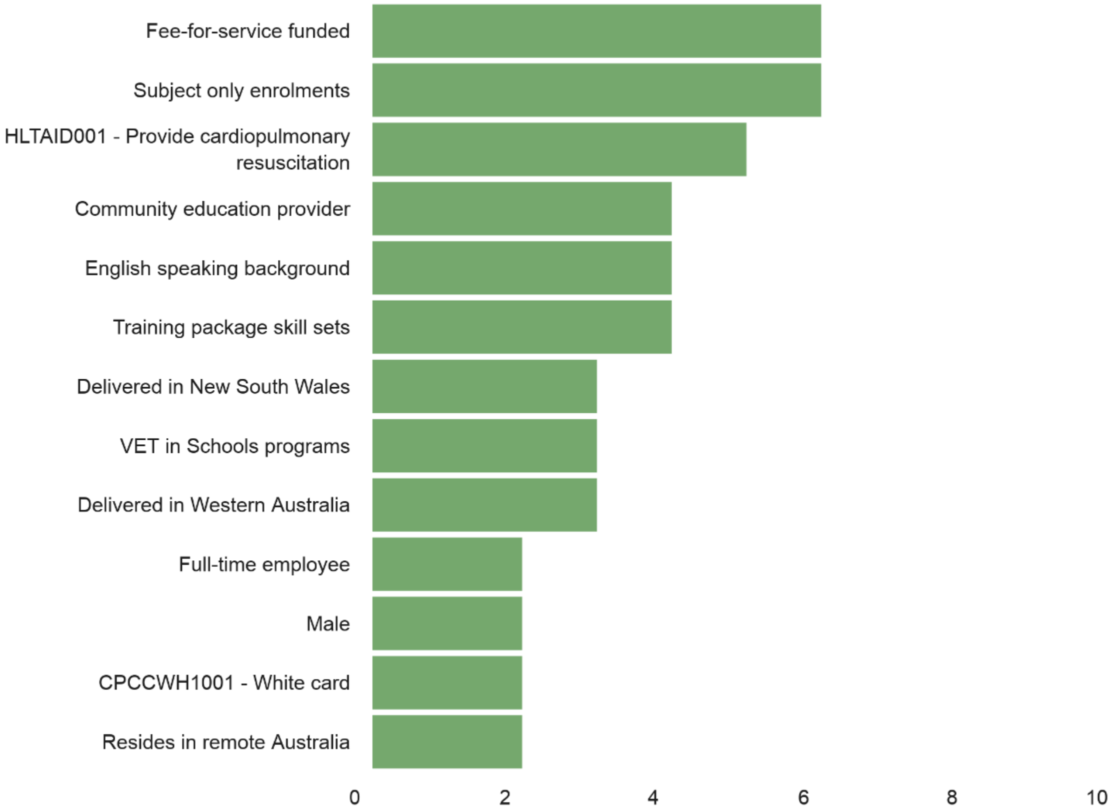
Several programs were seen disproportionately in one state or territory by comparison with the other states and territories. In some cases, it looked as though this was driven by state government policy priorities; in others, it resembled a market response to local economic factors and opportunities.

Figure 11 Jurisdictional priorities word cloud



Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 12 Jurisdictional priorities top features

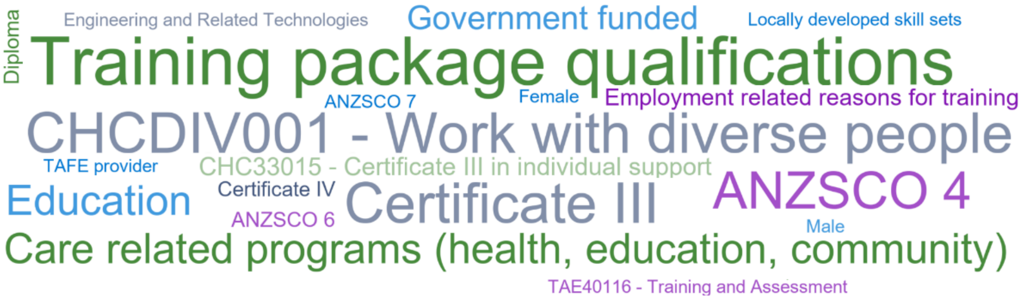


Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

Program enrolments NEI

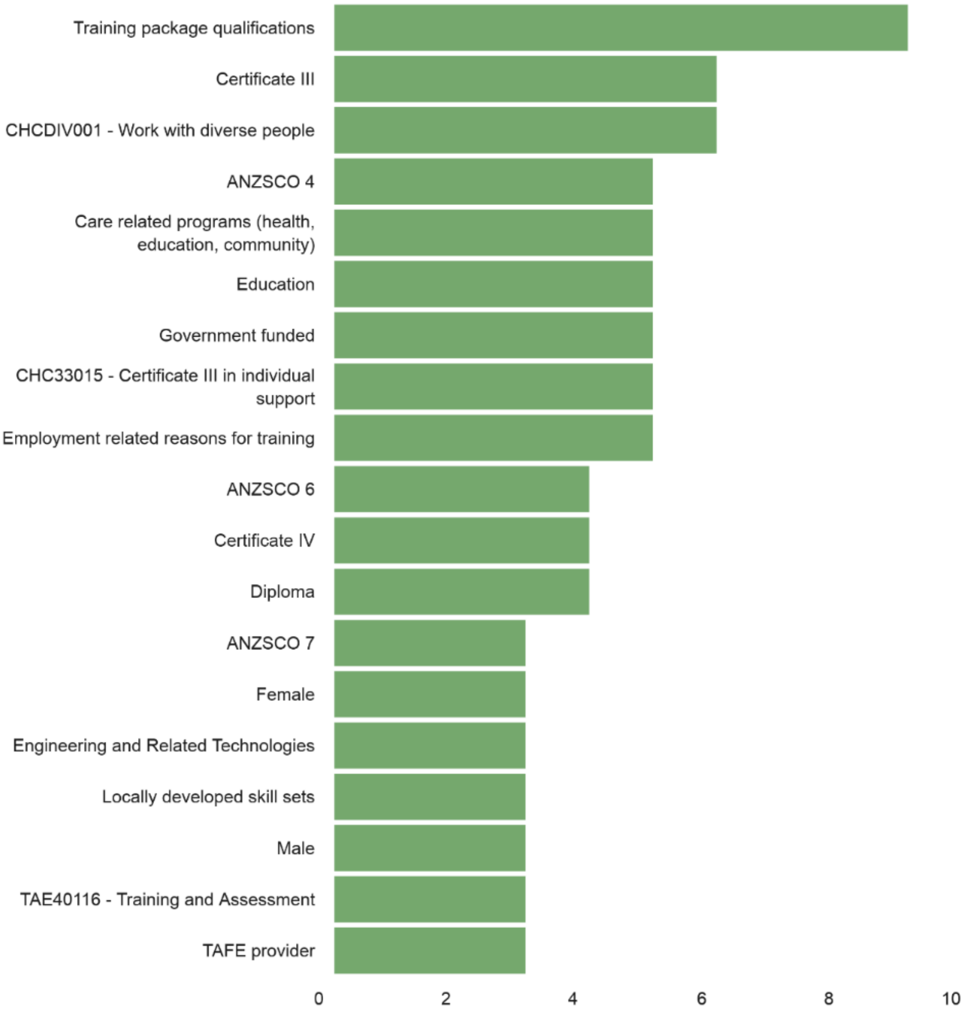
The last two segments capture those clusters not otherwise allocated above. The Program enrolments NEI segment captures clusters where at least 80% of the training was provided as part of a national training package qualification, an accredited qualification, an accredited course, a locally developed course, or a locally developed skill set.

Figure 13 Program enrolments NEI word cloud[4]



Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 14 Program enrolments NEI top features[4]



Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

[4] The ANZSCO classification is a code that uniquely identifies the type of occupation which a client would be qualified in on completion of their training. ANZSCO 4 refers to Community and Personal Service Workers, ANZSCO 6 to Sales Workers and ANZSCO 7 to Machinery Operators and Drivers.

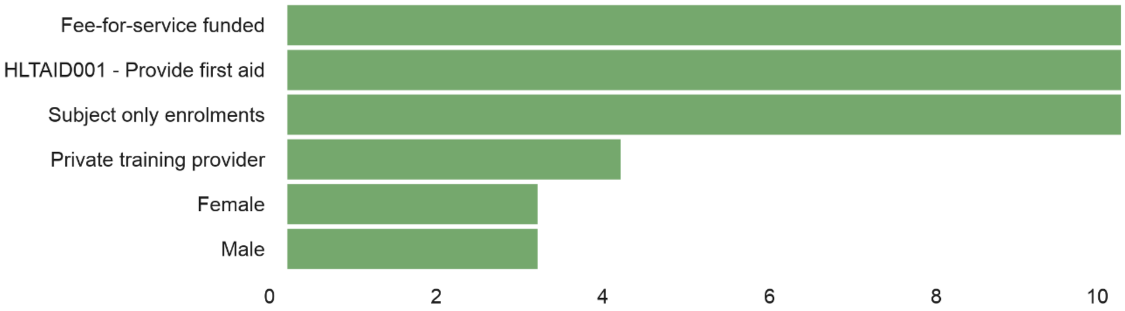
Subject only enrolments NEI

The Subject only enrolments NEI segment captured those clusters where at least 80% of the students were enrolled in stand-alone subjects. The most popular unit of competency in this segment was HLTAID001 - Provide cardiopulmonary resuscitation. Training in this segment is typically provided on a fee-for-service basis.

Figure 15 Subject only enrolments NEI word cloud

Note: The above includes features that were identified in two or more clusters allocated to this segment.

Figure 16 Subject only enrolments NEI top features



Note: The above shows the frequency in which a feature was identified across the clusters allocated to this segment. Data are presented where the frequency is greater than or equal to two.

DBSCAN

While k-means and agglomerative clustering proved useful, this was not the case with DBSCAN. A range of values were explored for epsilon (a maximum distance for a data point to be associated with another data point) and minimum samples (the minimum number of data points required to be within the epsilon radius for a point to be defined as a 'core point'). However, either the data did not cluster (most of the data points were identified as noise) or they clustered into largely one group, with noise as the second largest cluster (and often further clusters that were very small in size when compared with the first cluster and the unclustered noise).

It appears the DBSCAN algorithm was not well suited to the 2019 TVA data, and this approach was not further considered. This should not be seen as a reflection on the DBSCAN algorithm, merely its utility in this specific context.

A closer examination of three segments

Why go back to the original data?

We return to the original data to better understand the segments identified through the cluster analysis for several reasons. First, a few student records were missing data items. In some cases, we removed those records; in other cases we made assumptions about the missing data items. However, the absence of records means the cluster proportions are not necessarily reflective of the population proportions.

Second, the clustering algorithms assume that individuals belong to one, and only one, cluster. While this is a useful analytical simplification, it becomes problematic when clusters are not so clearly delineated in practice. In this case, we expect the three clusters selected to have some overlaps in membership.

Finally, we can better focus on a particular cohort of students using the original data than we can when attributing features to the clusters generated by the clustering algorithm.

Three selected market segments

The selection process began with a base population of students who were:

- doing one of the following: a training package qualification, an accredited qualification, an accredited course, a locally developed course
- not in or at school
- not an overseas student studying in Australia.

The size of the base population was 1.8 million students.

The three selected market segments we will explore are:

- migrants — people born overseas
- social inclusion — people with limited prior education and/or other markers of potential disadvantage
- targeted English — people undertaking specific English education programs.

Using text-pattern matching, we selected people undertaking targeted English programs. The program name had to include one of the following words: ‘English’, ‘literacy’, ‘in eal’ (where EAL stands for English as an additional language), or ‘general education for adults’. However, the program name must also exclude the following words: ‘diploma’, ‘advance’, ‘teach’, ‘health literacy’, ‘TAE’, ‘ESL’, ‘assessment and training’, ‘address’ and ‘early language and literacy’. The exclusions were to cover higher-level English programs and people who were training to teach English as a second language. Almost 95 000 students were selected in this segment. It should be noted that this designation of ‘targeted English’ is much narrower than the common VET term ‘foundation skills’.

We selected migrants based on their country of birth (other than Australia). There were 492 000 students in this segment.

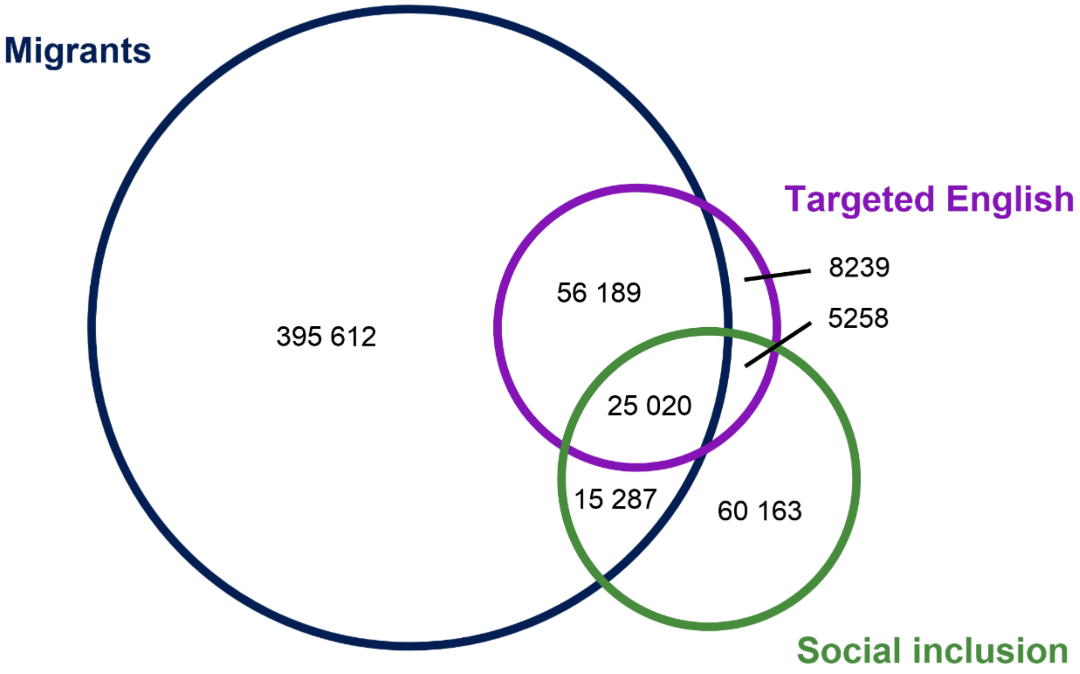
The social inclusion cohort was selected as having not completed Year 12, undertaking a program at certificate II level or lower, and displaying at least one of the following characteristics:

- living in the bottom 40% of SEIFA locations (using the Socio-Economic Indexes for Areas developed by the Australian Bureau of Statistics)
- having a disability
- being Indigenous
- having a primary language other than English.

There were almost 106 000 students in this segment.

The overlapping relationship between these segments is illustrated in the following weighted Venn diagram.

Figure 17 Illustration of segment overlap



Of note is the degree of overlap between the three segments. Most of the enrolment activity undertaken by migrants did not fall within the targeted English or social inclusion segments. However, the majority of the targeted English segment and just under half of the social inclusion segment were migrants.

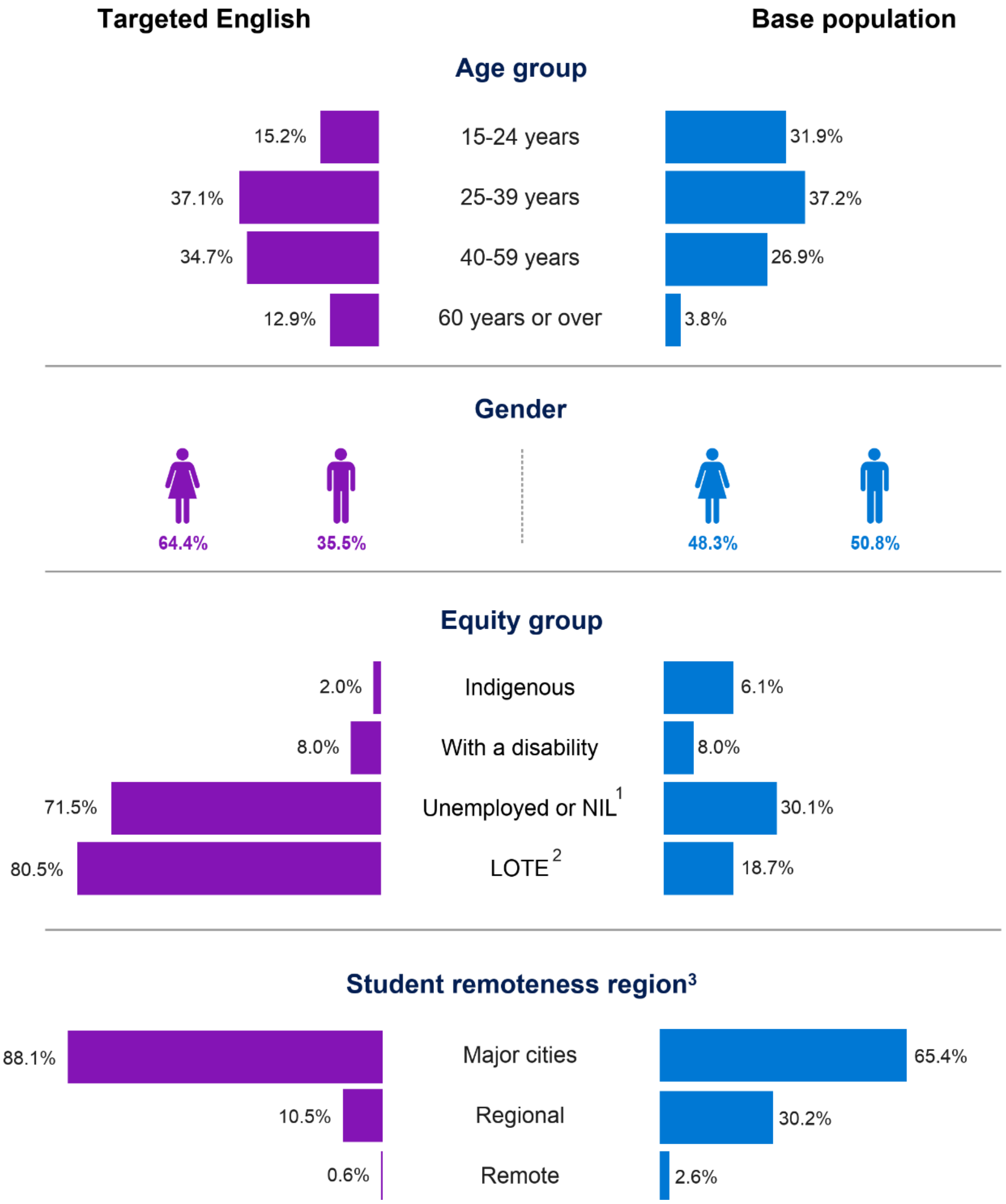
By means of case studies, the following section provides a deeper investigation into the student and training characteristics for each segment. It compares each segment with the base population[5] to highlight what differentiates them from each other and from VET students more broadly.

[5] The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia. There were 1.8 million students in the base population.

Case study: Targeted English

Compared with the base population, the targeted English segment was more likely to comprise of females, students aged 40 years and those who were unemployed or not in the labour force. They were also more likely to speak a language other than English as the main language at home, and reside in major cities, when compared with the base population.

Figure 18 Targeted English selected student characteristics



Note: Percentages in the above graphic will not sum to 100% as both ‘Other’ and ‘Not known’ responses are not presented here.

The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

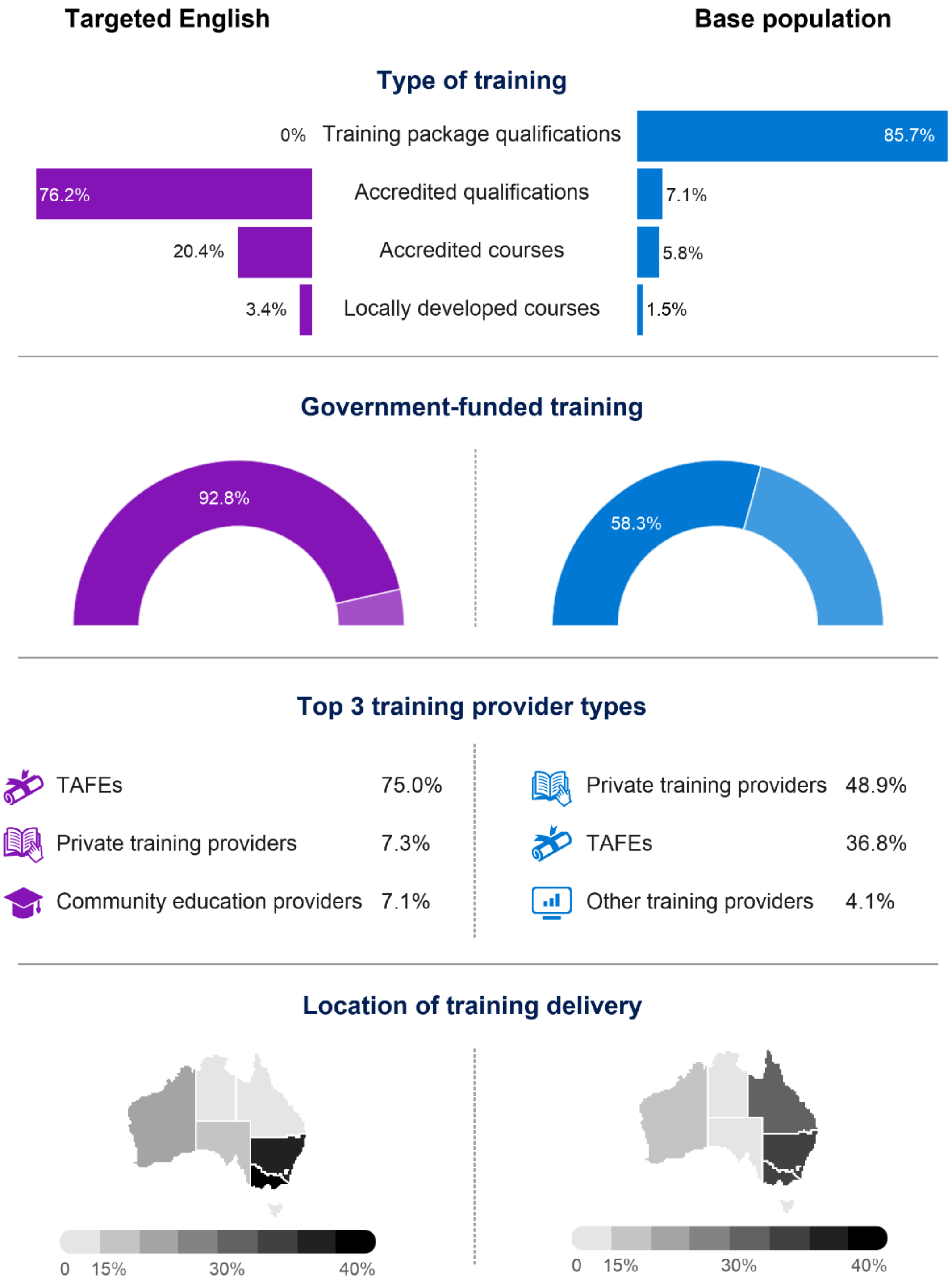
[1] NIL represents ‘Not in the labour force’.

[2] LOTE represents ‘Language other than English spoken at home’.

[3] Student remoteness region is based on the ARIA+ classification where remoteness is described in terms of the ease or difficulty residents face in accessing services.

Students in this segment were predominantly enrolled in government-funded training (92.8%), comprising of accredited qualifications and courses. They were more likely than the base population to undertake training at a TAFE provider and to have received training in New South Wales and Victoria. They were less likely than the base population to have received training in Queensland.

Figure 19 Targeted English selected training characteristics



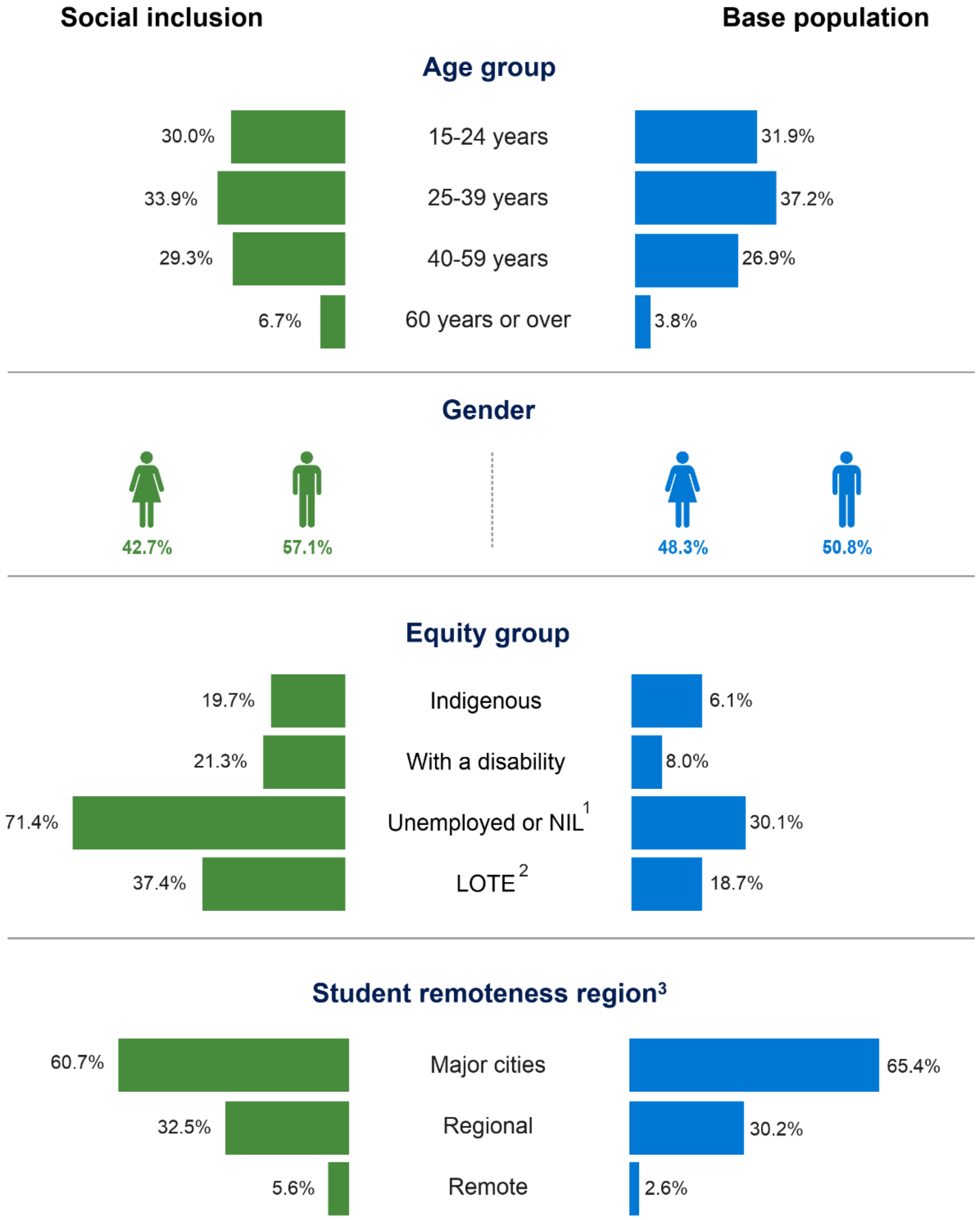
Note: Percentages in the above graphic will not sum to 100% as both 'Other' and 'Not known' responses are not presented here.

The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

Case study: Social inclusion

Compared with the base population, the social inclusion segment was more likely to comprise of males and students belonging to an equity group. They were also slightly more likely to be residing in a regional or remote area.

Figure 20 Social inclusion selected student characteristics



Note: Percentages in the above graphic will not sum to 100% as both ‘Other’ and ‘Not known’ responses are not presented here.

The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

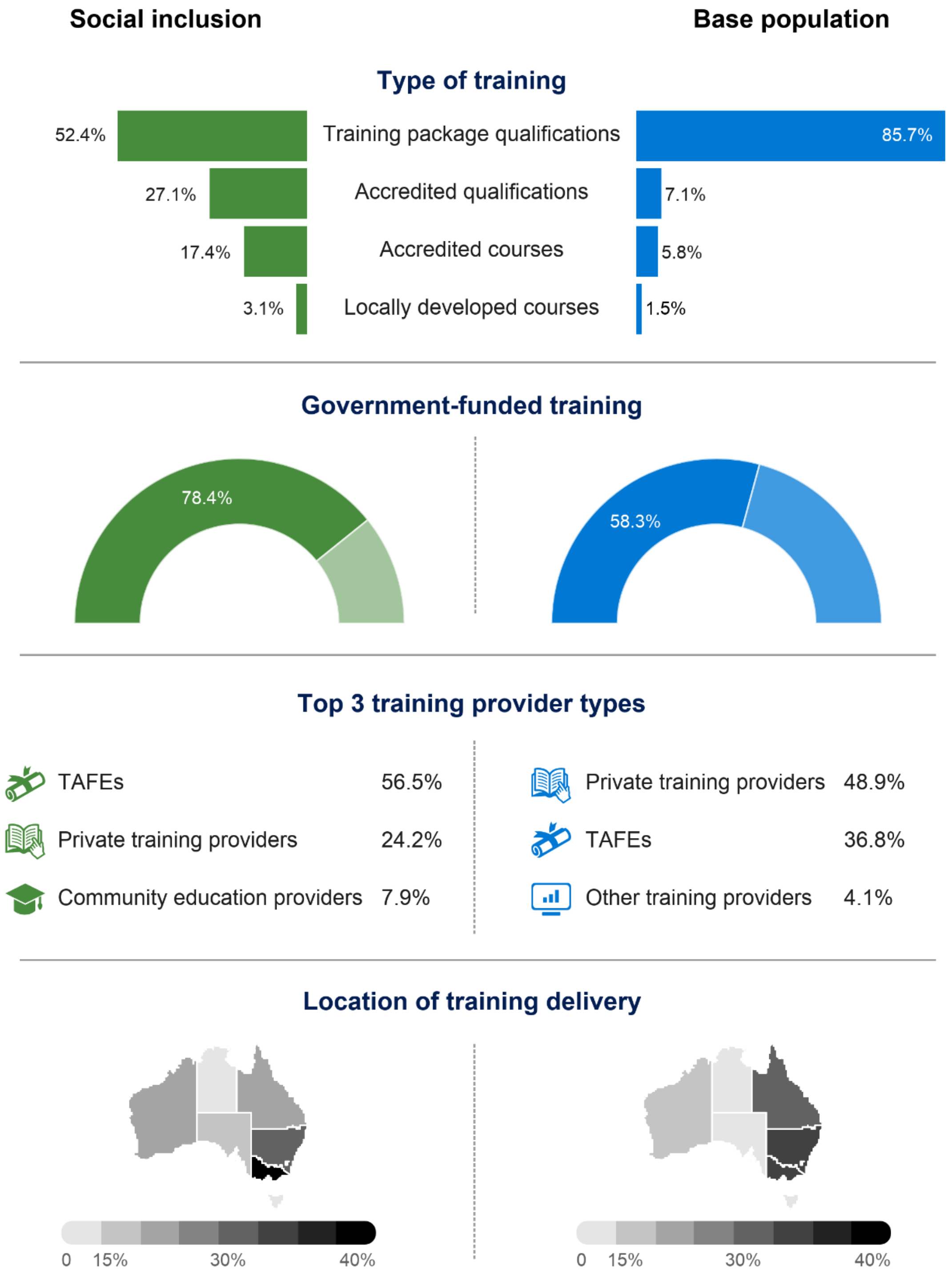
[1] NIL represents ‘Not in the labour force’.

[2] LOTE represents ‘Language other than English spoken at home’.

[3] Student remoteness region is based on the ARIA+ classification where remoteness is described in terms of the ease or difficulty residents face in accessing services.

Over three quarters of students in this segment were enrolled in government-funded training. When compared with the base population, training was less likely to comprise of training package qualifications, with a higher proportion of students enrolled in an accredited qualification or course and with a TAFE provider. Training was more likely to be delivered in Victoria and less likely to be delivered in Queensland, when compared with the base population.

Figure 21 Social inclusion selected training characteristics

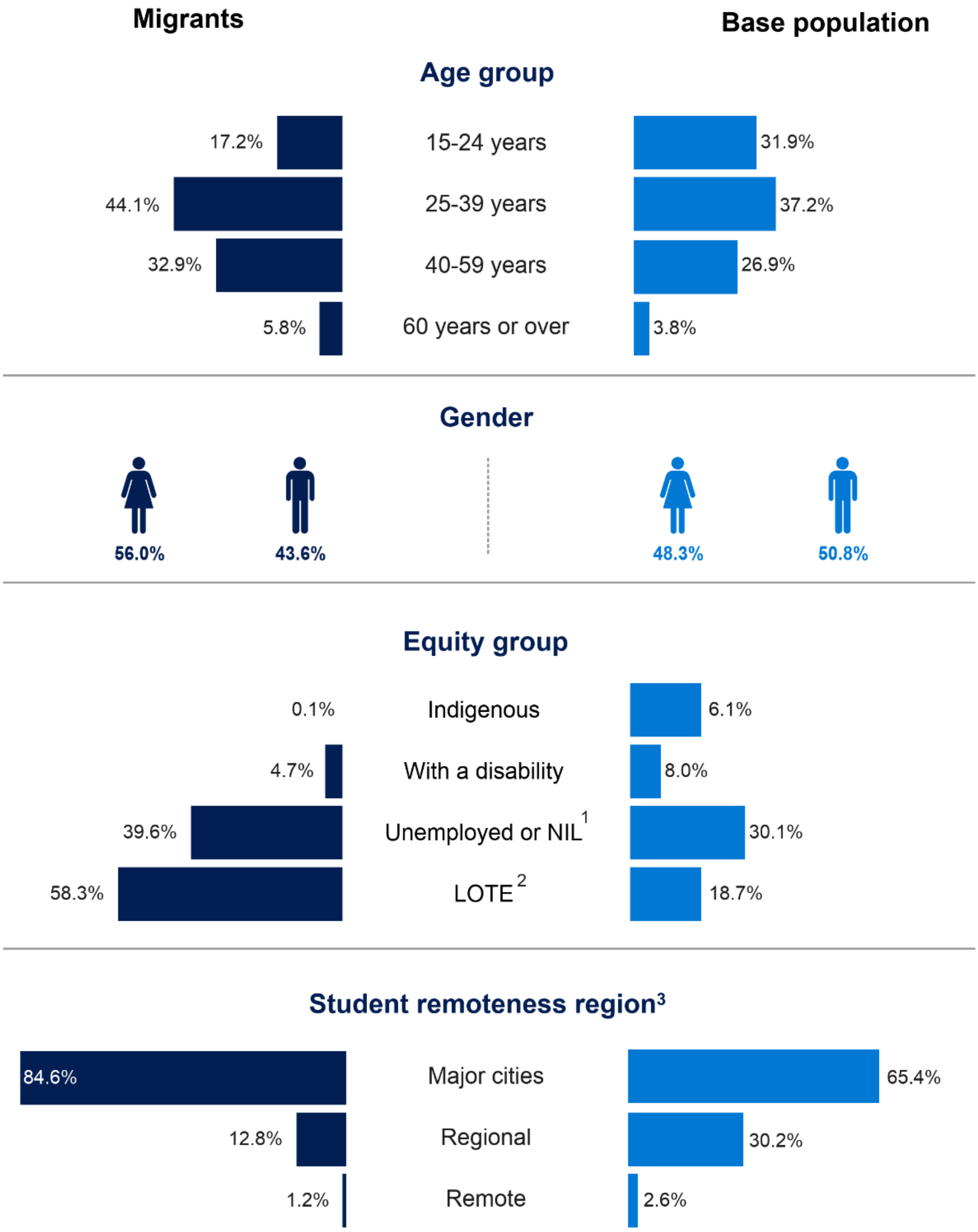


Note: Percentages in the above graphic will not sum to 100% as both ‘Other’ and ‘Not known’ responses are not presented here.
The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

Case study: Migrants

The migrants segment was more likely than the base population to comprise of females and students aged 25 to 59 years. It consisted of a higher proportion of students that were unemployed or not in the labour force and for which a language other than English was the main language spoken at home. Students were also more likely than the base population to reside in major cities.

Figure 22 Migrants selected student characteristics



Note: Percentages in the above graphic will not sum to 100% as both 'Other' and 'Not known' responses are not presented here.

The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

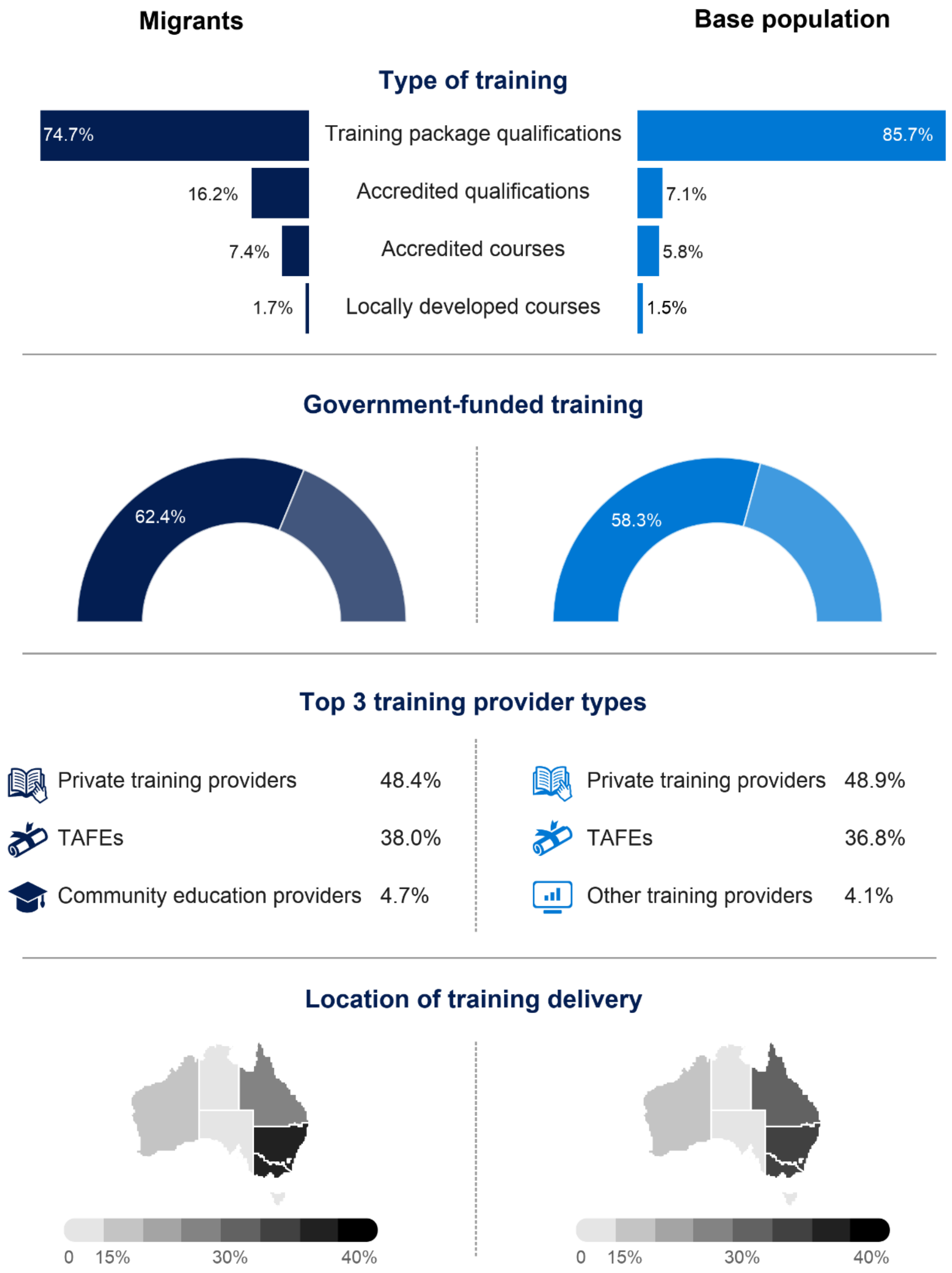
[1] NIL represents 'Not in the labour force'.

[2] LOTE represents 'Language other than English spoken at home'.

[3] Student remoteness region is based on the ARIA+ classification where remoteness is described in terms of the ease or difficulty residents face in accessing services.

This segment was less likely than the base population to undertake a training package qualification, with a higher proportion enrolled in an accredited qualification or course. Community education providers played a more important role for these students than for the base population.

Figure 23 Migrants selected training characteristics



Note: Percentages in the above graphic will not sum to 100% as both ‘Other’ and ‘Not known’ responses are not presented here.
The base population consists of students doing one of the following: a training package qualification, an accredited qualification, an accredited course, or a locally developed course; not in or at school; and not an overseas student studying in Australia.

The above case studies help to better understand the differences between the segments and how the student and training characteristics compare with the base population. By mapping segment features back to 2019 TVA data, other segments could be further explored in a similar manner.

References

Burkov, A 2019, *The hundred page machine learning book*, Burkov, Quebec City, Canada.

Circelli, M & Stanwick, J 2020, *Initial and continuing VET in Australia: clarifying definitions and applications*, NCVER, Adelaide.

Geron, A 2019, *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow*, O'Reilly Media, Sebastopol, CA.

Gower, JC 1971, 'A general coefficient of similarity and some of its properties', *Biometrics*, vol.27, no.4, December, 1971, pp.857—71.

Marsland, S 2015, *Machine learning: an algorithmic perspective*, 2nd edn, CRC Press, Boca Raton, FLA.

Moodie, G, Wheelahan, L, Fredman, N & Bexley, E 2015, *Towards a new approach to mid-level qualifications*, NCVER, Adelaide.

Müller, AC & Guido, S 2017, *Introduction to machine learning with Python: a guide for*

data scientists, O'Reilly Media, Sebastopol, CA.

NCVER (National Centre for Vocational Education research) 2020, *Total VET students and courses 2019*, NCVER, Adelaide.

——2021, *Total VET students and courses 2020,* NCVER, Adelaide.

Palmer, B 2021, *An analysis of 'micro-credentials' in VET*, NCVER, Adelaide.

Pedregosa, F et al. 2001, 'Scikit-learn: machine learning in Python', *Journal of Machine Learning*, vol.12, pp.2825—30.

Raschka, S & Mirjalili, V 2019, *Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*, 3rd edn, Packt, Birmingham, UK.

Statistics and Machine Learning in R undated, <https://github.com/Statistics-and-Machine-Learning-with-R/Statistical-Methods-and-Machine-Learning-in-R/wiki/Clustering>, accessed on 4 November 2021.

Appendix A – Dataset preparation for analysis

The 2019 TVA dataset contains largely categorical variables, with different levels of consistency and completeness, for millions of students, programs, and subjects. As a result, it is not well suited to the application of clustering algorithms and it required substantial treatment before it could be used. This appendix describes the preparation process.

The analysis commenced with an investigation of the TVA 2019 subject enrolment records relating to all types of VET learning: training package qualifications, accredited qualifications and courses, as well as single-subject enrolments that were not part of a recognised program of study. We were interested to see whether clusters emerged that included students studying recognised qualifications, as well as bundles of subjects as single-subject enrolments. There were some 27.5 million subject records in the 2019 TVA dataset.

From subjects to students

From these subject records, 5.7 million programs of learning were constructed. These comprised the unique students enrolled in a program at a specific registered training organisation (RTO). All the subject enrolments for a student without a recognised program at an RTO were also grouped together as a subject enrolment-based program of learning. It is possible for individual students to have two or more programs of learning; for example, if a student was enrolled at two different RTOs or if a student undertook a recognised program and some single-subject enrolments.

Because program codes are being continuously updated, older program codes were superseded to reflect the most recent code.

In the process of bringing the subject records together into a program of learning, anomalies identified in the values that pertained to the student (rather than the subject) were treated as follows:

- For student age, the mean age from the subject records was used (except age 0, which indicated missing data).
- The reported hours for each subject were summed.
- The level of education for the program and the highest level of education achieved by the student were encoded as an ordinal variable and averaged.
- The apprenticeship flag was set if it was set against any of the subjects.
- For the categorical variables, the mode (most frequently occurring) value was selected. However, if there was more than one modal value, the data were marked as missing.

Data transformations

Each of the three algorithms expect continuous numerical data that have been similarly scaled. The algorithms work best if the data are well distributed across the domain of its scale. None of the algorithms accept records with missing data. The algorithms do not accept categorical data.

As noted earlier, in its raw form, the 2019 TVA data are not well suited for clustering. Most of the data items in TVA are categorical data and most of them have at least some records with missing data. Some of the data items have up to 50% of records with missing data. A small number of data items contained data that are 'bunched together'. For example, the ages of students range from 9 to 99 years, but the interquartile range (the middle 50% of) is 22 to 44 years.

Missing values need to be treated; categorical values need to be transformed; skewed distributions need to be corrected.

Missing data treatments

Since the clustering algorithms cannot handle missing data, several standard treatments for missing data have been developed, all of which were used in this report:

- record removal — remove the records with missing data
- attribution — attribute values where data are missing
- encoded inclusion — encode missing data as a separate category, or exclude missing categorical values when they are converted to numerical values via one hot encoding
- data item exclusion — exclude the data item with missing values from the analysis.

Where less than 1% of the data items for a variable were missing, the affected student records were removed from the dataset for analysis. This included the student's age (0.4% of records had missing age data); the delivery mode (0.8%); the national funding source (0.3%); the VET in Schools flag (0.2%); the program type of training (0.1%); and the state of delivery location (0.6%).

Some missing data were treated by attribution. If the primary language information was missing, it was coded as English, but only if the country of birth was Australia. If the disability flag was missing, it was coded to not with a disability. If the Indigenous identifier was missing, it was coded as non-Indigenous. If the at-school flag was missing, it was coded as not at school.

Categorical variables with moderate amounts of missing data were left in the dataset. However, when the categorical items were encoded, the missing items were not encoded. This issue will be discussed further below, under the treatment of categorical data with one hot encoding. Variables affected in this way include gender (3% of the data items were missing); labour force status (23%); remoteness (7%); SEIFA (12%); country of birth (11.7%); and language (12%).

Finally, several data items were not included in the analysis because the items had too many missing values. For example, not included in the analysis were: the program field of education (54% missing); the ANZSCO classification for a program (59%); and the reason for study (44%).

Categorical data treatments

As noted above, the level of education and the highest education level were changed from categorical variables to ordinal variables on the scale shown in table A1:

Table A1 Categorical data treatments

Category	Value
Not known	0
Did not go to school	1
Year 9 or lower	2
Miscellaneous education	3
Certificate I	4
Year 10	5
Certificate II	6
Year 11	7
Year 12	8
Certificate III	9
Certificate IV	10
Diploma	11
Advanced diploma/associate degree	12
Bachelor degree or above	13

The remaining categorical variables were converted to numerical data using one hot encoding. For each categorical variable, a column of data was created for each of the possible values of the categorical item. For example, with the gender variable, three data columns were created: male, female and non-binary. In each of these columns, a one (1) was encoded if the original variable contains the value for that column. Otherwise, the column was coded as zero (0). Continuing with our example, male was encoded into the three columns as [1, 0, 0]. Female was encoded as [0, 1, 0]. Non-binary was encoded as [0, 0, 1]. And missing data were encoded as [0, 0, 0].

Because one hot encoding adds to the dimensionality of the data, and as high-dimensional data can be more challenging analytically than low-dimensional data, a couple of further adjustments were made. First, some data items were simplified. For example, the language spoken at home was simplified to English or other; the country of birth was simplified to Australia or other. Second, where a data item was effectively a binary item (with perhaps some missing data), only the analytically more interesting of two encoded columns was retained. For example, only the language - other, and country of birth - other columns were brought into the analysis.

Feature scaling

Because most of the features in the data were categorical items on the scale from zero through to one, the same scale was applied to the numerical data items. Before the analysis, all numerical items were transformed to be on the scale from zero to one.

Other treatments

Prior to feature scaling, some data items were subject to a logarithmic transformation to better distribute right-skewed data items more evenly (in a frequency distribution, a long tail is on the right side of the x-axis for right-skewed data). The items to which this technique was applied include student age (capped to exclude those aged over 65 years); the number of subjects studied (capped to exclude over 15 subjects); and the number of study hours reported (capped to exclude over 1000 hours).

Programs of learning where the delivery location was overseas were excluded from the dataset before the analysis.

The completion of the data-preparation steps outlined above resulted in 5.5 million student records, which were then taken into the analysis phase.

The Gower distance matrix

To this point we have discussed the data preparation process that was applied to the k-means algorithm. With the agglomerative and DBSCAN clustering, we precomputed the 'distances' between each of the students and provided the precomputed distance matrix to the clustering algorithm. Furthermore, given that the size of the distance matrix grows by a factor of N2 for N students, we limited the analysis to a random sample of 60 000 students (to ensure that it was computationally tractable).

For the precomputed distance matrix, a Gower (1971) dissimilarity matrix was constructed. This allowed the calculation of distances based on a combination of numerical and categorical data. The process works in the following way:

- For each numerical feature, we create a NxN scaled dissimilarity matrix, where all the distances between students are between 0 and 1. Items of the same value are coded as 0 and the most different items become 1.
- For each categorical feature, we have a binary dissimilarity matrix, where sameness is coded as 0, and differences are coded as 1.

We then summarise these dissimilarity matrices for each feature in the data by taking their arithmetic mean. This will yield a single matrix with every element in the scale from zero to one.

The traditional Gower matrix was augmented to include a dissimilarity matrix for the subjects studied. First, a subject similarity matrix was calculated. For each pair of students, we divided the number of subjects that both individuals studied in common (the number of subjects in the intersection set) by the total number of unique subjects studied (the number of subjects in the union set for the student pair). Second, to get the dissimilarity matrix, we subtracted each element of the similarity matrix from 1 (that is, dissimilarity = 1 - similarity). This subject dissimilarity matrix was then included in the step of taking the arithmetic mean over all of the individual feature dissimilarity matrices.

It should be noted that the Gower dissimilarity matrix was calculated with reference to the original categorical variables rather than the one hot encoded variables. It also should be noted that missing categorical data in the sample were treated as their own category.

Final caveats

The 2019 TVA dataset is comprised of administrative data. There are artefacts in the data that arise from different administrative practices in the states and territories, while different administrative practices arise in different educational settings. For example, more missing data are associated with single-subject enrolments by comparison with nationally recognised programs. These artefacts can become the structures in the data that the clustering algorithms detect and cluster against.

While steps were taken to minimise the impact of missing data and other artefacts that might have arisen from administrative practices, it is worth being aware of the potential impact of administrative practices when considering the output from the clustering algorithms.

Appendix B – Identified features and market segments

The list of clusters and the associated market segments for each of the four clustering approaches is set out in the following tables. In these tables the age range relates to the 25th, 50th and 75th percentiles. The first word in each list of features is a simplified version of the variable name, followed by the value for the variable. Some of these variable names need explanation: ProgType is Program Type; LOE is level of education; FOE is field of education; Funding is the funding source; LF is labour force status; Org is the type of organisation providing the training; Lang is the primary language spoken at home; and Apps is an apprenticeship or traineeship.

Table B1 K-means with 8 clusters

	N	Student %	Age range	Features	Segment label	Possible segment labels
0	1289624	23.32	[25.0, 35.0, 47.0]	[PopularSubject HLTAID001 31.0%, !ProgType stand-alone subject enrolments 94.7%, !Funding fee-for-service 98.1%, !Gender Male 97.5%]	Subject only enrolments NEI	Subject only enrolments NEI
1	643988	11.65	[20.0, 28.0, 40.0]	[PopularSubject CHCDIV001 4.2%, ProgType Training package qualifications 62.3%, ProgType Accredited qualifications 16.4%, ProgType Locally developed skill sets 11.3%, ProgType Accredited courses 6.1%, ProgType Locally developed courses 1.9%, !Funding government funding 90.1%, LF Not employed - not seeking employment 21.4%, FOE Mixed Field Programmes 22.1%, FOE Agriculture, Environmental and Related Studies 3.4%, FOE Information Technology 3.1%, Apps True 19.9%, !Org TAFE 90.8%, LOE Miscellaneous education 20.2%, LOE Certificate I 9.1%, ANZSCO 3 30.3%, TargEng TE 14.2%]	Social inclusion	Social inclusion, Program enrolments NEI
2	834713	15.1	[25.0, 35.0, 47.0]	[PopularSubject HLTAID001 65.5%, !ProgType stand-alone subject enrolments 96.3%, !Funding fee-for-service 99.5%, !Org Private training provider 93.0%]	Subject only enrolments NEI	Subject only enrolments NEI
3	269692	4.88	[23.0, 27.0, 31.0]	[PopularSubject BSBMGT517 8.6%, PopularProgram BSB50420 9.4%, !ProgType Training package qualifications 95.4%, !Funding fee-for-service 99.4%, !Birth Other 94.3%, Lang Other 75.2%, FOE Management and Commerce 51.6%, FOE Food, Hospitality and Personal Services 11.6%, FOE Information Technology 4.8%, Org Private training provider 89.3%, LOE Diploma 35.7%, LOE Advanced	Overseas students	Overseas students, Migrants, Program enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
				diploma/Associate degree 13.4%, Remote Overseas 89.6%, ANZSCO 1 24.8%, ANZSCO 5 23.4%, ANZSCO 2 10.1%]		
4	698439	12.63	[22.0, 31.0, 43.0]	[PopularSubject CHCDIV001 18.3%, PopularProgram CHC33015 7.2%, !ProgType Training package qualifications 95.4%, Funding government funding 54.5%, !Gender Female 99.3%, FOE Society and Culture 30.4%, FOE Education 13.4%, StdyRsn employment related 51.3%, LOE Certificate III 41.3%, LOE Certificate IV 26.9%, LOE Diploma 19.1%, ANZSCO 4 49.1%, ANZSCO 6 4.9%, Care Program 19.9%]	Program enrolments NEI	Program enrolments NEI
5	873067	15.79	[25.0, 35.0, 48.0]	[PopularSubject HLTAID001 55.7%, !ProgType stand-alone subject enrolments 92.5%, !Funding fee-for-service 97.3%, !Gender Female 97.3%]	Subject only enrolments NEI	Subject only enrolments NEI
6	355737	6.43	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject BSBWOR203 12.4%, PopularProgram SIT20316 6.2%, !ProgType Training package qualifications 90.2%, Funding government funding 81.2%, FOE Food, Hospitality and Personal Services 16.0%, FOE Creative Arts 5.6%, FOE Information Technology 4.4%, FOE Agriculture, Environmental and Related Studies 3.5%, Org School 39.4%, LOE Certificate II 61.0%, LOE Certificate I 11.8%, ANZSCO 8 25.6%, !School Yes 97.2%]	Younger students	Younger students, Program enrolments NEI
7	564034	10.2	[23.0, 31.0, 43.0]	[PopularSubject BSBWOR301 4.6%, PopularProgram TAE40116 4.8%, !ProgType Training package qualifications 96.3%, !Gender Male 98.8%, FOE Engineering and Related Technologies 31.9%, FOE Architecture and Building 11.8%, Apps True 17.7%, LOE Certificate III 47.7%, ANZSCO 7 15.6%]	Program enrolments NEI	Program enrolments NEI

Table B2 K-means with 16 clusters

	N	Student %	Age range	Features	Segment label	Possible segment labels
0	75260	1.36	[27.0, 37.0, 49.0]	[PopularSubject HLTAID001 93.3%, !ProgType stand-alone subject enrolments 98.3%, !Funding fee-for-service 99.8%, !State Western Australia 99.7%, !Lang English 90.8%, LF Full-time employee 64.4%, !Org Community education provider 99.8%, !School Yes 92.5%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI
1	174903	3.16	[26.0, 33.0, 42.0]	[PopularSubject HLTAID001 28.6%, !ProgType stand-alone subject enrolments 91.1%, ProgType Locally developed courses 2.1%, !Funding fee-for-service 96.8%, !Birth Other 98.5%, !Lang Other 99.2%, Org Private training provider 87.1%, Remote Major cities of Australia 87.2%]	Migrants	Migrants, Subject only enrolments NEI
2	311236	5.63	[24.0, 32.0, 45.0]	[PopularSubject CPCCWHS1001 18.6%, ProgType stand-alone subject enrolments 89.9%, ProgType Training package skill sets 4.1%, !Funding fee-for-service 96.6%, !State New South Wales 100.0%, !Gender Male 96.7%, !Org Private training provider 95.5%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI
3	476130	8.61	[23.0, 32.0, 43.0]	[PopularSubject CHCDIV001 18.0%, PopularProgram CHC33015 7.8%, !ProgType Training package qualifications 97.5%, !Gender Female 99.1%, FOE Society and Culture 32.5%, FOE Education 13.9%, StdyRsn employment related 49.9%, !Org Private training provider 91.4%, LOE Certificate III 43.2%, LOE Certificate IV 26.9%, LOE Diploma 18.3%, ANZSCO 4 50.8%, ANZSCO 6 5.5%, Care Care Program 21.1%]	Program enrolments NEI	Program enrolments NEI
4	569479	10.3	[24.0, 34.0, 47.0]	[PopularSubject HLTAID001 46.5%, !ProgType stand-alone subject enrolments 93.0%, !Funding fee-for-service 98.8%, !Gender Female 98.9%, Org Private training provider 87.1%]	Subject only enrolments NEI	Subject only enrolments NEI
5	411225	7.44	[25.0, 34.0, 47.0]	[PopularSubject HLTAID001 60.1%, !ProgType stand-alone subject enrolments 95.8%, !Funding fee-for-service 99.1%, Gender Male 79.6%, !Org Private training provider 91.3%]	Subject only enrolments NEI	Subject only enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
6	248658	4.5	[23.0, 26.0, 31.0]	[PopularSubject BSBMGT517 9.0%, PopularProgram BSB50420 9.8%, !ProgType Training package qualifications 96.4%, !Funding fee-for-service 99.7%, !Birth Other 94.0%, Lang Other 74.1%, FOE Management and Commerce 53.6%, FOE Food, Hospitality and Personal Services 12.2%, FOE Information Technology 5.1%, Org Private training provider 89.3%, LOE Diploma 37.2%, LOE Advanced diploma/Associate degree 14.0%, !Remote Overseas 95.8%, ANZSCO 1 26.2%, ANZSCO 5 24.0%]	Overseas students	Overseas students, Migrants, Program enrolments NEI
7	398444	7.21	[20.0, 25.0, 36.0]	[PopularSubject CPCCOHS2001A 4.6%, ProgType Training package qualifications 87.7%, ProgType Locally developed skill sets 8.5%, Funding government funding 89.8%, !Gender Male 99.9%, Lang English 86.1%, FOE Engineering and Related Technologies 31.2%, FOE Architecture and Building 17.2%, FOE Agriculture, Environmental and Related Studies 5.3%, FOE Information Technology 4.1%, Apps True 36.2%, Org TAFE 85.0%, Org University 6.7%, LOE Certificate III 46.7%, ANZSCO 3 47.4%]	Program enrolments NEI	Program enrolments NEI
8	362920	6.56	[21.0, 30.0, 42.0]	[PopularSubject CHCDIV001 16.7%, ProgType Training package qualifications 87.3%, ProgType Locally developed skill sets 9.1%, Funding government funding 87.7%, !Gender Female 99.8%, FOE Health 11.8%, FOE Education 10.1%, FOE Creative Arts 3.6%, StdyRsn employment related 52.2%, StdyRsn further study 4.9%, Org TAFE 83.0%, Org University 5.3%, LOE Diploma 18.6%, ANZSCO 4 40.6%, Care Care Program 14.5%]	Program enrolments NEI	Program enrolments NEI
9	345115	6.24	[27.0, 37.0, 49.0]	[PopularSubject HLTAID001 27.3%, !ProgType stand-alone subject enrolments 91.1%, !Funding fee-for-service 97.0%, State Queensland 50.7%, State Tasmania 4.6%, State Northern Territory 3.8%, !Gender Male 94.6%, LF Full-time employee 56.9%, Remote Inner regional Australia 50.1%, Remote Outer regional Australia 31.5%, Remote Remote Australia 5.9%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
10	417152	7.54	[25.0, 35.0, 48.0]	[PopularSubject HLTAID001 70.6%, !ProgType stand-alone subject enrolments 94.5%, !Funding fee-for-service 99.0%, !Gender Female 99.8%, StdyRsn job requirement 45.9%, !Org Private training provider 94.0%]	Subject only enrolments NEI	Subject only enrolments NEI
11	429634	7.77	[26.0, 34.0, 46.0]	[PopularSubject HLTAID001 26.4%, !ProgType stand-alone subject enrolments 93.1%, !Funding fee-for-service 98.8%, !Gender Male 97.9%, LF Full-time employee 55.9%, !Remote Major cities of Australia 99.9%]	Subject only enrolments NEI	Subject only enrolments NEI
12	341346	6.17	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject BSBWOR203 12.6%, PopularProgram SIT20316 6.4%, !ProgType Training package qualifications 91.3%, Funding government funding 82.2%, FOE Food, Hospitality and Personal Services 16.2%, FOE Creative Arts 5.8%, FOE Information Technology 4.5%, FOE Agriculture, Environmental and Related Studies 3.4%, Org School 40.8%, LOE Certificate II 61.9%, LOE Certificate I 12.0%, ANZSCO 8 26.4%, !School Yes 97.7%]	Younger students	Younger students, Program enrolments NEI
13	501339	9.07	[23.0, 32.0, 43.0]	[PopularSubject BSBWOR301 5.1%, PopularProgram TAE40116 4.8%, !ProgType Training package qualifications 99.1%, !Gender Male 98.7%, FOE Engineering and Related Technologies 32.3%, Org Private training provider 89.1%, LOE Certificate III 48.8%, ANZSCO 7 16.6%, ANZSCO 6 4.2%]	Program enrolments NEI	Program enrolments NEI
14	299416	5.42	[27.0, 39.0, 51.0]	[Age Older, PopularSubject HLTAID001 82.5%, !ProgType stand-alone subject enrolments 95.5%, !Funding fee-for-service 95.2%, State New South Wales 61.3%, !Org Community education provider 99.2%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI
15	167037	3.02	[24.0, 35.0, 48.0]	[PopularSubject SWERWT001 4.2%, ProgType Accredited qualifications 67.0%, ProgType Accredited courses 23.9%, ProgType Locally developed courses 1.9%, Funding government funding 84.9%, Birth Other 71.8%, Lang Other 65.3%, LF Not employed - not seeking employment 39.9%,	Targeted English	Targeted English, Migrants, Social inclusion, Program enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
				FOE Mixed Field Programmes 72.7%, StdyRsn personal reasons 24.2%, StdyRsn further study 8.2%, Org TAFE 72.5%, Org University 6.1%, LOE Miscellaneous education 29.3%, LOE Certificate I 24.0%, Remote Major cities of Australia 83.4%, ANZSCO 2 11.7%, TargEng TE 56.8%]		

Table B3 Agglomerative clustering with 8 clusters

	N	Student %	Age range	Features	Segment label	Possible segment labels
0	2868	4.78	[22.0, 26.0, 31.0]	[PopularSubject BSBMGT517 9.4%, PopularProgram BSB50420 9.8%, !ProgType Training package qualifications 91.9%, !Funding fee-for-service 97.5%, Birth Other 89.0%, Lang Other 67.9%, FOE Management and Commerce 50.1%, FOE Food, Hospitality and Personal Services 11.6%, FOE Information Technology 5.3%, Org Private training provider 85.7%, LOE Diploma 35.8%, LOE Advanced diploma/Associate degree 14.2%, !Remote Overseas 90.8%, ANZSCO 1 25.3%, ANZSCO 5 22.1%, ANZSCO 2 10.9%]	Overseas students	Overseas students, Migrants, Program enrolments NEI
1	31250	52.08	[25.0, 35.0, 47.0]	[PopularSubject HLTAID001 46.2%, !ProgType stand-alone subject enrolments 92.7%, !Funding fee-for-service 98.7%]	Subject only enrolments NEI	Subject only enrolments NEI
2	2448	4.08	[26.0, 36.0, 49.0]	[PopularSubject VU21800 8.4%, PopularProgram 22300VIC 7.0%, ProgType Accredited qualifications 49.0%, ProgType Accredited courses 43.8%, Funding government funding 61.4%, State Australian Capital Territory 5.9%, Birth Other 52.0%, Lang Other 46.1%, LF Not employed - not seeking employment 25.7%, FOE Mixed Field Programmes 54.6%, FOE Health 17.1%, StdyRsn further study 5.6%, Org TAFE 50.9%, LOE Miscellaneous education 44.9%, LOE Certificate I 17.4%, Remote Major cities of Australia 78.0%, ANZSCO 2 12.7%, TargEng TE 40.1%]	Migrants	Migrants, Social inclusion, Jurisdictional priorities, Program enrolments NEI
3	4177	6.96	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject BSBWHS201 12.2%, PopularProgram FSK20119 5.7%, ProgType Training package qualifications	Younger students	Younger students,

	N	Student %	Age range	Features	Segment label	Possible segment labels
				89.6%, Funding government funding 74.2%, FOE Food, Hospitality and Personal Services 13.2%, FOE Creative Arts 5.9%, FOE Information Technology 4.1%, Org School 35.6%, LOE Certificate II 60.1%, LOE Certificate I 11.6%, ANZSCO 8 23.7%, ANZSCO 5 19.0%, School Yes 90.0%]		Program enrolments NEI
4	13590	22.65	[24.0, 32.0, 43.0]	[PopularSubject CHCDIV001 13.3%, !ProgType Training package qualifications 93.5%, Funding government funding 58.0%, FOE Education 9.4%, StdyRsn employment related 49.3%, LOE Certificate IV 26.9%, ANZSCO 4 35.2%, Care Care Program 12.1%]	Program enrolments NEI	Program enrolments NEI
5	684	1.14	[26.0, 36.0, 47.0]	[PopularSubject CPCCWHS1001 12.1%, ProgType Locally developed skill sets 74.3%, ProgType Training package skill sets 11.4%, Funding government funding 56.3%, State New South Wales 55.6%, State Australian Capital Territory 7.5%, State Northern Territory 5.8%, Indig Indigenous 13.2%, Org TAFE 85.7%, LOE Miscellaneous education 86.4%, Remote Remote Australia 6.4%]	Social inclusion	Social inclusion, Jurisdictional priorities, Subject only enrolments NEI
6	3151	5.25	[19.0, 21.0, 27.0]	[Age Younger, PopularSubject CPCCOHS2001A 5.7%, PopularProgram UEE30820 10.8%, !ProgType Training package qualifications 100.0%, !Funding government funding 97.6%, Gender Male 73.6%, Lang English 89.1%, LF Full-time employee 72.6%, FOE Engineering and Related Technologies 38.6%, FOE Architecture and Building 19.5%, FOE Food, Hospitality and Personal Services 11.8%, FOE Agriculture, Environmental and Related Studies 3.6%, StdyRsn job requirement 48.2%, !Apps True 96.4%, Org TAFE 46.1%, LOE Certificate III 83.8%, ANZSCO 3 60.0%, ANZSCO 7 8.0%, ANZSCO 6 5.1%]	Younger students	Younger students, Program enrolments NEI
7	1832	3.05	[19.0, 30.0, 45.0]	[PopularSubject HLTAID001 79.0%, !ProgType stand-alone subject enrolments 90.1%, ProgType Training package skill sets 7.9%, !Funding fee-for-service 100.0%, State Western Australia 47.7%, Lang English 88.7%, Org Community education provider 49.6%, School Yes 71.6%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI

Table B4 Agglomerative clustering with 16 clusters

	N	Student %	Age range	Features	Segment label	Possible segment labels
0	6000	10.0	[21.0, 30.0, 42.0]	[PopularSubject CHCDIV001 14.4%, ProgType Training package qualifications 88.2%, ProgType Locally developed skill sets 9.3%, !Funding government funding 94.2%, Disabs Yes 15.8%, !Birth Australia 90.5%, !Lang English 90.3%, FOE Creative Arts 3.5%, FOE Agriculture, Environmental and Related Studies 3.2%, StdyRsn employment related 53.7%, Org TAFE 55.7%, ANZSCO 4 35.7%, Care Care Program 12.5%]	Program enrolments NEI	Program enrolments NEI
1	27533	45.89	[25.0, 35.0, 47.0]	[PopularSubject HLTAID001 44.4%, !ProgType stand-alone subject enrolments 93.0%, !Funding fee-for-service 98.6%]	Subject only enrolments NEI	Subject only enrolments NEI
2	1124	1.87	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject HLTAID003 13.4%, PopularProgram SIS30115 6.7%, ProgType Training package qualifications 83.6%, ProgType Accredited qualifications 15.8%, !Funding government funding 93.1%, State Victoria 63.7%, FOE Society and Culture 27.6%, FOE Creative Arts 6.8%, Org Enterprise provider 15.4%, Org School 12.0%, LOE Certificate II 60.7%, ANZSCO 4 32.5%, ANZSCO 8 14.4%, !School Yes 99.5%]	Younger students	Younger students, Jurisdictional priorities, Program enrolments NEI
3	1832	3.05	[19.0, 30.0, 45.0]	[PopularSubject HLTAID001 79.0%, !ProgType stand-alone subject enrolments 90.1%, ProgType Training package skill sets 7.9%, !Funding fee-for-service 100.0%, State Western Australia 47.7%, Lang English 88.7%, Org Community education provider 49.6%, School Yes 71.6%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI
4	1417	2.36	[28.0, 39.0, 51.0]	[Age Older, PopularSubject VU21800 14.5%, PopularProgram 22300VIC 12.1%, ProgType Accredited courses 74.2%, ProgType Accredited qualifications 14.3%, State Australian Capital Territory 9.4%, FOE Mixed Field Programmes 36.3%, FOE Health 28.8%, LOE Miscellaneous education 75.7%, ANZSCO 2 17.3%, TargEng TE 20.2%]	Social inclusion	Social inclusion, Jurisdictional priorities, Program enrolments NEI
5	5966	9.94	[26.0, 34.0, 45.0]	[PopularSubject CHCDIV001 10.4%, PopularProgram TAE40116 6.6%, !ProgType Training package qualifications 98.3%, FOE Management and Commerce 28.5%, FOE	Program enrolments NEI	Program enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
				Education 10.5%, LOE Certificate III 37.6%, LOE Certificate IV 31.9%, LOE Diploma 18.6%, ANZSCO 2 11.8%, ANZSCO 7 8.7%, ANZSCO 6 4.8%]		
6	3151	5.25	[19.0, 21.0, 27.0]	[Age Younger, PopularSubject CPCCOHS2001A 5.7%, PopularProgram UEE30820 10.8%, !ProgType Training package qualifications 100.0%, !Funding government funding 97.6%, Gender Male 73.6%, Lang English 89.1%, LF Full-time employee 72.6%, FOE Engineering and Related Technologies 38.6%, FOE Architecture and Building 19.5%, FOE Food, Hospitality and Personal Services 11.8%, FOE Agriculture, Environmental and Related Studies 3.6%, StdyRsn job requirement 48.2%, !Apps True 96.4%, Org TAFE 46.1%, LOE Certificate III 83.8%, ANZSCO 3 60.0%, ANZSCO 7 8.0%, ANZSCO 6 5.1%]	Younger students	Younger students, Program enrolments NEI
7	1043	1.74	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject HLTWHS001 15.8%, PopularProgram SIS30115 8.6%, ProgType Training package qualifications 83.3%, ProgType Accredited qualifications 12.6%, !Funding fee-for-service 94.0%, !Birth Australia 93.8%, Lang English 88.1%, LF Not employed - not seeking employment 34.5%, LF Unemployed - seeking part-time work 19.4%, FOE Society and Culture 29.6%, FOE Creative Arts 7.3%, StdyRsn personal reasons 24.4%, StdyRsn further study 8.3%, Org School 8.6%, LOE Certificate II 42.3%, LOE Certificate III 37.4%, LOE Certificate I 11.1%, ANZSCO 4 33.4%, ANZSCO 5 21.5%, School Yes 66.3%]	Younger students	Younger students, Social inclusion, Program enrolments NEI
8	160	0.27	[16.0, 17.0, 40.0]	[Age Younger, PopularSubject CPCCOHS2001A 23.1%, PopularProgram CPC10120 23.1%, !ProgType Training package qualifications 93.1%, FOE Architecture and Building 26.2%, FOE Food, Hospitality and Personal Services 11.2%, FOE Creative Arts 4.4%, Org Enterprise provider 10.0%, LOE Certificate II 35.0%, LOE Certificate I 23.8%, ANZSCO 8 35.6%, ANZSCO 5 20.6%, School Yes 68.8%]	Younger students	Younger students, Program enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
9	3717	6.19	[24.0, 33.0, 47.0]	[PopularSubject HLTAID001 59.9%, !ProgType stand-alone subject enrolments 90.6%, !Funding fee-for-service 99.9%]	Subject only enrolments NEI	Subject only enrolments NEI
10	2708	4.51	[22.0, 26.0, 31.0]	[PopularSubject BSBMGT517 9.9%, PopularProgram BSB50420 10.3%, !ProgType Training package qualifications 91.9%, !Funding fee-for-service 98.0%, !Birth Other 93.5%, Lang Other 71.6%, FOE Management and Commerce 51.8%, FOE Food, Hospitality and Personal Services 11.6%, FOE Information Technology 5.6%, Org Private training provider 86.3%, LOE Diploma 37.5%, LOE Advanced diploma/Associate degree 15.0%, !Remote Overseas 96.0%, ANZSCO 1 26.7%, ANZSCO 5 22.2%, ANZSCO 2 11.3%]	Overseas students	Overseas students, Migrants, Program enrolments NEI
11	1031	1.72	[23.0, 33.0, 44.0]	[PopularSubject SWERWT001 6.4%, PopularProgram 10727NAT 8.1%, !ProgType Accredited qualifications 96.8%, !Funding government funding 97.8%, Disabs Yes 15.0%, Birth Other 66.7%, Lang Other 62.8%, LF Not employed - not seeking employment 35.8%, FOE Mixed Field Programmes 79.8%, StdyRsn personal reasons 25.6%, StdyRsn further study 9.4%, Org TAFE 73.8%, Org University 8.4%, LOE Certificate I 34.8%, LOE Certificate II 30.9%, Remote Major cities of Australia 83.1%, TargEng TE 67.5%]	Targeted English	Targeted English, Migrants, Social inclusion, Program enrolments NEI
12	161	0.27	[29.0, 38.0, 46.0]	[PopularSubject CPCCWHS1001 6.2%, ProgType Locally developed skill sets 46.0%, ProgType Training package skill sets 16.1%, Funding government funding 72.7%, State New South Wales 58.4%, State Northern Territory 24.8%, Gender Female 64.6%, Indig Indigenous 24.8%, Birth Other 72.0%, Lang Other 86.3%, LF Unemployed - seeking part-time work 24.2%, Org TAFE 55.3%, LOE Miscellaneous education 62.1%]	Migrants	Migrants, Social inclusion, Jurisdictional priorities, Mixed enrolments NEI
13	1624	2.71	[26.0, 35.0, 43.0]	[PopularSubject CHCDIV001 20.3%, PopularProgram CHC33015 8.8%, !ProgType Training package qualifications 95.8%, !Funding government funding 92.5%, Gender Female 65.2%, Birth Other 85.5%, Lang Other 64.2%, LF Unemployed - seeking part-time work 17.9%, FOE Society and Culture 30.0%,	Migrants	Migrants, Social inclusion, Program enrolments NEI

	N	Student %	Age range	Features	Segment label	Possible segment labels
				FOE Education 11.6%, StdyRsn employment related 61.0%, Org TAFE 52.3%, LOE Certificate IV 26.9%, LOE Diploma 21.2%, !Remote Major cities of Australia 90.1%, ANZSCO 4 48.0%, Care Care Program 23.2%]		
14	523	0.87	[24.5, 36.0, 47.0]	[PopularSubject CPCCWHS1001 14.0%, ProgType Locally developed skill sets 83.0%, ProgType Training package skill sets 9.9%, Funding government funding 51.2%, State New South Wales 54.7%, State Australian Capital Territory 8.6%, Lang English 89.3%, !Org TAFE 95.0%, !LOE Miscellaneous education 93.9%, Remote Remote Australia 7.3%]	Jurisdictional priorities	Jurisdictional priorities, Subject only enrolments NEI
15	2010	3.35	[16.0, 16.0, 17.0]	[Age Younger, PopularSubject BSBWOR203 15.6%, PopularProgram FSK20119 9.4%, !ProgType Training package qualifications 96.1%, !Funding government funding 99.0%, State Queensland 49.4%, !Lang English 93.6%, FOE Food, Hospitality and Personal Services 17.5%, FOE Architecture and Building 12.3%, FOE Information Technology 6.3%, FOE Creative Arts 4.8%, FOE Agriculture, Environmental and Related Studies 4.0%, Org School 62.7%, LOE Certificate II 69.1%, LOE Certificate I 15.6%, ANZSCO 8 34.7%, !School Yes 97.0%]	Younger students	Younger students, Jurisdictional priorities, Program enrolments NEI

NCVER

National Centre for Vocational Education Research

Level 5, 60 Light Square, Adelaide, SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

Phone +61 8 8230 8400 **Email** ncver@ncver.edu.au
Web <https://www.ncver.edu.au> <https://www.lsay.edu.au>
Follow us: <https://twitter.com/ncver> <https://www.linkedin.com/company/ncver>