MONASH University

Final Report

Disha Rathod

Report for
Data Intelligence and Insights Department of Monash

26 October 2024

MONASH
BUSINESS
SCHOOL

Department of Econometrics & Business Statistics

(04) 0328 1394
drat0009@student.monash.edu

ABN: 12 377 614 012

1 Introduction

2 Background, Motivation

The project is driven by a clear and a very impactful goal where the Australian government has set a target for universities to ensure that at least 20% of their undergraduate students come from Low Socio-Economic Status (SES) backgrounds. This target aims to promote equality in terms of access to higher education, giving students from disadvantaged communities a fairer chance at achieving a university education. However, achieving this goal is not straightforward, it requires a detailed understanding of the current demographics, regional socioeconomic disparities, and the barriers that students face for which we have used SEIFA methodology.

The primary focus of this project is the state of Victoria, which has regions with diverse socio-economic backgrounds. We aim to analyze the feasibility of meeting the government's target by forecasting the Year 12 graduates population which are the students who are potential university candidates—over the coming years. To do this, we're using data from the Australian Bureau of Statistics (ABS), specifically the 2021 Census, which provides insights into age demographics and socio-economic status across Victoria.

The analysis isn't just about numbers; it's about understanding the real challenges that students from Low SES backgrounds face. For example, some of the barriers include geographical distance to campuses, financial pressures that might force students to work rather than study, and differences in access to educational resources. The project will use R programming to visualize these factors on Victoria's map, highlighting where populations are concentrated and identifying areas where students may have fewer opportunities to access higher education.

One key aspect of this project is understanding how the population and socio-economic landscape will evolve. Using statistical tools, we aim to predict changes in the Year 12 graduates population from 2021 to 2030. By combining this demographic forecasting with SES data, we can identify regions that may need targeted support to meet the enrollment goals.

Beyond the analysis, the project has a broader purpose: to provide actionable insights for Monash University. By identifying patterns and trends, we can inform strategies to make higher education more accessible. This might involve expanding campuses closer to disadvantaged areas, offering free travel passes to students from Low SES regions, or designing flexible academic schedules to accommodate those who need to balance work and study. The project's findings could directly shape university policies and government decisions, ensuring that support is directed where it's most needed.

In summary, this project isn't just about hitting enrollment targets—it's about understanding and overcoming the challenges that disadvantaged students face in accessing education. It's a step towards a fairer and more inclusive education system, where all students, regardless of background, have a chance to succeed.

3 Objectives and Significance

3.1 Key Objectives

Our project centered on understanding and forecasting Year 12-ready populations in Victoria from 2021 to 2030, with a focus on socio-economic disparities. By projecting future student numbers, we aimed to identify which regions would experience growth or decline, allowing Monash University to make informed decisions about resource allocation. These projections help Monash anticipate changes in enrollment demands and prepare accordingly, from expanding campus facilities to focusing recruitment and retention efforts where they are needed most.

To get a clear picture of socio-economic challenges, we used the Socio-Economic Indexes for Areas (SEIFA) methodology, which ranks regions based on various indicators like income, education, and employment. This analysis helped us pinpoint the most disadvantaged areas, highlighting where socio-economic barriers could hinder access to education. Understanding these disparities allows Monash University to target its outreach and support programs more effectively, ensuring resources go to regions with the greatest need.

We combined statistical analysis with geospatial mapping to create a comprehensive view of the educational landscape in Victoria. While statistical data gave us the numbers, the maps allowed us to visualize how these figures played out geographically. This combination not only showed us what was happening but also why certain regions faced greater challenges. It provided a clearer context for Monash, helping them understand the socio-economic dynamics at play and informing where to focus efforts for the greatest impact.

The ultimate goal was to provide Monash University with data-driven recommendations to increase Low SES enrollments. By merging population projections with socio-economic analysis, we highlighted key areas for intervention. This guidance helps the university make strategic decisions about where to direct scholarships, outreach, and academic support to ensure it aligns with their mission to improve access for disadvantaged students. Our approach, grounded in both numbers and real-world context, offers a pathway for Monash to foster a more equitable and inclusive educational environment across Victoria.

3.2 Significance of the Analysis and Methodology

Our project aimed to support equitable access to education by using SEIFA to identify socio-economic disparities in Victoria. This analysis helped Monash University pinpoint areas facing the most significant challenges, ensuring their initiatives are grounded in evidence and targeted for maximum impact. The demographic and socio-economic analysis offered a clear roadmap for strategic resource allocation, whether it's expanding facilities in high-growth areas or providing additional support where student numbers are declining. By combining statistical projections with geospatial mapping, we were able to see the human context behind the data, showing how socio-economic conditions intersect with geography. This comprehensive approach provided Monash with valuable insights to address the real challenges faced by students, making their outreach and support efforts more effective. Overall, the project delivered a data-driven view of Victoria's Year 12 landscape, identifying both areas of need and opportunities for growth. This blend of demographic analysis, SEIFA insights, and visual techniques equips Monash with practical recommendations to boost Low SES enrollment and foster a more inclusive education environment.

4 Methodology

4.1 Introduction to Methodology

The goal of this project is to help Monash University meet the government's target of increasing Low SES student participation (i.e. 20%). Over the past 12 weeks, I was working at Monash's Intelligence and Analytics unit, led by Patrick Leung, has been dedicated to this analysis, using a range of data tools and visualization techniques in R.

Our approach began with collecting and preparing data from the 2021 Australian Bureau of Statistics (ABS) Census, which provided crucial information on the Year 12 graduates population and socio-economic indicators. We used R to clean and merge these datasets, managing the challenges of incomplete and unavailable information. The focus was on ensuring that the data was accurate and reliable for further analysis.

The heart of our methodology was projecting how the Year 12 population in Victoria would change from 2021 to 2030, with a particular emphasis on differences across regions with varying SES profiles. R was instrumental in this, with tools like tidyverse helped us handle data, while sf and ggplot2 allowed us to create detailed maps. These visualizations were crucial for spotting regional disparities and identifying areas that might need more support.

Overall, our methodology was about more than just crunching numbers; it was about using data to tell a story and inform strategies for making higher education more accessible and equitable. This approach is designed to support Monash in planning for a fairer education system, ensuring that every student, regardless of background, has a chance to succeed.

4.2 Data Collection and Preparation

###Data Sources

The backbone of this project was the data provided by the 2021 Census from the Australian Bureau of Statistics (ABS). This data was invaluable in helping me understand the population distribution and socio-economic factors in Victoria, particularly focusing on students who are likely to be Year 12 graduates in the coming years. By looking at the age demographics and socio-economic indicators from the census, we could identify which areas were likely to have higher or lower socio-economic advantages. SEIFA methodology was the key that helpes us in defining these SES status.

We used data from three geographical levels: Statistical Area 1 (SA1), Statistical Area 3 (SA3) and Statistical Area 4 (SA4). SA1 data gave us a detailed, view, allowing us to pinpoint local trends and disparities, while the broader SA3 and SA4 levels helped us capture regional patterns. This multi-level approach provided a clear picture of how socio-economic conditions varied across the state, identifying both the most disadvantaged areas and those with greater opportunities.

To make the data more manageable and specific to our needs, we used the ABS Table Builder, a powerful tool that allowed us to create custom datasets. Table Builder helped us focus on the Year 12-ready age group and the socio-economic factors that are relevant to understanding educational access. This customization was crucial because the ABS data covers a vast range of topics, and we needed to narrow it down to the most relevant information for our analysis.

We also gathered additional datasets to enrich our analysis, including data for socio-economic rankings and the locations of university campuses across Victoria. Incorporating campus locations allowed us to examine how access to higher education varies by region, helping us understand the impact of distance and proximity on Low SES students.

###Data Preprocessing

Preparing the data for analysis was a critical step in ensuring that our results would be accurate and meaningful. We undertook a thorough data preparation process, which involved several key tasks:

```
[1] 6336925
```

```
# A tibble: 10 x 2
```

Table 1: *Total Values Summarized by Year*

Year	Total_Value
2021	70594.33
2022	70604.33
2023	71845.33
2024	74063.67
2025	76009.33
2026	76796.33
2027	77068.33
2028	77195.00
2029	77943.00
2030	78787.33

```
   Year  Total_Value
   <chr>       <dbl>
 1 2021       70594.
 2 2022       70604.
 3 2023       71845.
 4 2024       74064.
 5 2025       76009.
 6 2026       76796.
 7 2027       77068.
 8 2028       77195.
 9 2029       77943.
10 2030       78787.
```

```
tibble [190 x 3] (S3: tbl_df/tbl/data.frame)
 $ SA4_NAME_2021: chr [1:190] "Ballarat" "Ballarat" "Ballarat" "Ballarat" ...
 $ Year         : num [1:190] 2021 2022 2023 2024 2025 ...
 $ Total_Value  : num [1:190] 1897 1949 2066 2138 2198 ...
```

```
Rows: 59,280
Columns: 5
$ X2021.Statistical.Area.Level.1..SA1. <dbl> 10102100701, 10102100702, 1010210~
$ State                                <chr> "NSW", "NSW", "NSW", "NSW", "NSW"~
$ Usual.Resident.Population            <int> 305, 301, 471, 522, 423, 290, 416~
$ Score                                <dbl> 984.3059, 1072.3003, 970.2893, 97~
$ Percentile.within.State              <int> 40, 68, 35, 36, 43, 28, 46, 57, 6~
```

```
Rows: 15,014
Columns: 5
$ SA1reg                    <dbl> 20101100101, 20101100102, 20101100105, 20101~
$ State                     <chr> "VIC", "VIC", "VIC", "VIC", "VIC", "VIC", "V~
$ Usual.Resident.Population <int> 435, 184, 377, 584, 358, 791, 527, 513, 366,~
$ Score                     <dbl> 939.4502, 993.0258, 882.7877, 951.0035, 852.~
$ Percentile.within.State   <int> 22, 40, 9, 25, 5, 36, 15, 69, 61, 48, 54, 38~

Rows: 150,140
Columns: 9
$ SA1reg                    <dbl> 20101100101, 20101100101, 20101100101, 20101~
$ Year                      <chr> "2021", "2022", "2023", "2024", "2025", "202~
$ Value                     <dbl> 9.333333, 7.333333, 7.000000, 6.666667, 6.66~
$ State                     <chr> "VIC", "VIC", "VIC", "VIC", "VIC", "VIC", "V~
$ Usual.Resident.Population <int> 435, 435, 435, 435, 435, 435, 435, 435, 435,~
$ Score                     <dbl> 939.4502, 939.4502, 939.4502, 939.4502, 939.~
$ Percentile.within.State   <int> 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 40, ~
$ SA4_NAME_2021             <chr> "Ballarat", "Ballarat", "Ballarat", "Ballara~
$ Percentile_Category       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2,~
```

Reading and Merging Datasets: - The first step was to bring all the relevant datasets into R using the `tidyverse` package. This included demographic data from the ABS Census, the SES data and the Statistical Area data. Each dataset had its own structure, with some in CSV format and others in Excel. We used functions like `read.csv()` and `read_excel()` to load the data into data frames in R. - Merging these datasets was a significant task. Since the data came from different sources, each had its own unique identifiers and formats. We used `left_join()` and `inner_join()` from the `dplyr` package to match records using geographical codes like SA1, SA3 and SA4 identifiers. This merging allowed us to integrate the three datasets into a single, comprehensive dataset.

Handling Missing Data: - Missing data is a common challenge in any analysis, and our project was no exception. Some areas, especially those with really low population,had to be removed using the `filter()` function. This was particularly important in ensuring that the analysis was not skewed by unreliable data. - Also, some of the data had gap with the first variable having multiple values in the second variable. To address this, we used the `fill()` function in `tidyverse`. This method ensured that missing values were logically estimated based on nearby data, maintaining the continuity of trends.

Formatting Age Categories for Projections: - A key focus of our analysis was to project the Year 12 graduates population from 2021 to 2030. This required a precise organization of age data, as we needed to account for how each age group would transition year by year. To do this, we grouped age data into three-year segments (e.g., 17, 18, and 19-year-olds for 2021) and calculated averages to make predictions for 2021, and used rolling averages for the following years. - Example: Years 16,17 and 18 were taken into consideration for 2022 Year 12 graduate forecast from the 2021 census data and 15, 16 and 17 were taken for 2023. - Using R's `mutate()` and `case_when()` functions, we created new columns that captured these transitions over time. This allowed us to project how the demographics would shift year by year, giving us a clear forecast of the Year 12 population. By structuring the data in this way, we were able to make consistent and reliable projections across the decade.

Data Cleaning and Standardization: - Data from different sources often comes in varying formats, so a thorough cleaning was necessary. This involved standardizing column names, aligning geographical codes, and converting text-based age groups into numerical categories using `as.numeric()`. Ensuring consistency was crucial for performing accurate calculations and comparisons. -We addressed discrepancies in region names by using `rename()` to standardize labels across datasets, ensuring consistency during merging and analysis.

Geospatial Data Preparation: - Since a large part of our analysis focused on mapping socio-economic trends, preparing the geospatial data was a major task. We used the `sf` package to handle spatial data and imported shapefiles that contained the geographical boundaries for Victoria's SA1 and SA3 regions. These shapefiles allowed us to link the data directly to specific locations. - Spatial joins were conducted to attach demographic information to these geographical areas, enabling us to create visual representations of the data. We used coordinate transformations to ensure that the spatial data aligned perfectly with the demographic datasets, allowing us to produce accurate and detailed maps.

Challenges in Data Preparation: - One of the biggest challenges we faced was aligning data from multiple sources. The ABS Census, SES indicators, and campus data were collected at different times and with slightly different methodologies. This meant that certain definitions and boundaries needed to be reconciled. For example, one dataset had more SA1 regions, which we had to remove when joining with the SES data, because those extra SA1 regions has incomplete data or very low population. - Merging data from various geographical levels (SA1 and SA3) presented another challenge. The detailed SA1 data is excellent for pinpointing local trends, but coverage can be inconsistent in sparsely populated areas. So, we had to aggregate data to the SA3 level to get a clearer picture of the underlying trends. This aggregation sometimes required recalculating averages and ensuring that the results remained accurate.

Through careful data collection, meticulous preparation, and a focus on accuracy, we created a robust dataset that served as the foundation for our analysis. This allowed us to confidently move forward with forecasting, visualizing, and drawing insights, knowing that our data was both reliable and comprehensive.

4.3 Forecasting Methodology

We projected the Year 12-ready population in Victoria from 2021 to 2030 to support Monash University's Low SES enrollment goals using demographic data and R.

Population Projection Approach

To estimate the future Year 12 graduate population, we used age data from the 2021 Census and employed a simple averaging method. For each forecast year, we averaged three age groups to get a reliable projection. For example, to determine the Year 12 graduate population in 2021, we looked at 17, 18, and 19-year-olds. For 2022, we averaged 16, 17, and 18-year-olds, and continued this pattern up to 2030. This technique helped us smooth out any year-to-year variations, ensuring a more stable and accurate projection.

The decision to use three-year averages was intentional, it allowed us to account for natural fluctuations in age groups without overemphasizing any particular year's anomalies. This method balances detail with simplicity, making it easier to track long-term trends without getting lost in minor yearly changes.

How R Was Used for Forecasting

R was the backbone of our statistical analysis, and its flexibility allowed us to handle complex data manipulation with ease. We relied heavily on R's data wrangling capabilities to prepare and forecast the Year 12-ready population. Here's how we approached the task:

Structuring Data for Analysis: - We used the `tidyverse` package in R to import and clean demographic data, using `mutate()` and `case_when()` to create projection columns for each year. - For example, the `case_when()` function allowed us to categorize and modify data based on specific conditions. This was crucial for shifting the focus from one age group to another as we moved through each forecast year.

Managing Regional Differences: - Since our analysis focused on various regions within Victoria, we grouped the data by geographical areas such as Statistical Area 1 (SA1) and Statistical Area 3 (SA3). This ensured that our forecasts reflected local trends, not just general state-wide averages. Grouping data by region allowed us to identify which areas might meet enrollment targets and which might need additional support.

Assumptions and Limitations

As with any forecasting project, our approach relied on several key assumptions that helped guide the analysis but also highlighted certain limitations:

Stable Demographic Trends: - We assumed that the factors like birth rates and migration to stay consistent without significant disruptions. While this assumption is generally safe for short-term projections, unexpected events—like economic changes or new policies—could alter these patterns.

Three-Year Averaging: - The decision to average across three age groups was made to smooth out any yearly fluctuations in the population data. However, this method also means that any sudden demographic changes, like a spike or drop in birth rates, might not be fully captured in the projections.

Cohort Stability: - Our projections assumed that students would generally progress through the education system in a consistent manner. This does not account for unexpected factors like changes in dropout rates or new education policies, which could affect the accuracy of our predictions.

Data Quality: - The accuracy of our forecasts depends heavily on the quality of the 2021 Census data. We took steps to clean and validate the data, but any inconsistencies or gaps in the original dataset could impact our results. Additionally, socio-economic indicators can evolve over time, which could affect long-term predictions if local dynamics shift significantly.

By acknowledging these assumptions, we aimed to maintain a transparent and realistic view of what our forecasting could and couldn't achieve. Despite the challenges, the use of R enabled us to adapt and make adjustments as needed, allowing for a flexible approach to demographic forecasting. This methodology gave us a solid foundation to predict future trends and support Monash's mission to make higher education more inclusive and accessible for students from all backgrounds.

4.4 Geospatial Data Preparation

Mapping socio-economic trends across Victoria was a crucial part of our analysis, as it provided a clear picture of regional differences in population and socio-economic status. This step was essential to understand which areas had the highest concentration of Year 12 graduate students and how access to Monash's campus varied by location. To achieve this, we needed to work with geospatial data, ensuring it was accurately aligned with our demographic and socio-economic information.

Using the `sf` Package and Shapefiles

We used the `sf` package in R to handle spatial data, starting with shapefiles that defined Victoria's Statistical Areas (SA1 for neighborhoods and SA3 for larger regions). We imported these into R with `st_read()`, ensuring the geographic data was accurate, as any errors could affect the entire analysis.

Merging Spatial and Demographic Data

Once we had the geographic boundaries set, we needed to connect them with the demographic data and SES data. This was done through a process called spatial joining, which links data to specific areas based on their location on the map. Using the `st_join()` function in R, we merged the census data with the boundaries of SA1 and SA3 regions. This allowed us to visualize things like socio-economic categories and population densities directly on the map.

One of the trickiest parts was making sure that all the datasets used the same coordinate system. Sometimes, different data sources use different ways to define locations, so we had to standardize everything to a common system using `st_transform()`. This step ensured that all the data matched up properly, preventing any errors in how regions were displayed on the maps.

Creating Maps and Visualizations

With the spatial data ready and merged with our demographic information, we could start creating visualizations. We used `ggplot2`, a powerful plotting tool in R, to generate maps that highlighted key trends across Victoria:

Socio-Economic Status (SES) Categories: - We mapped the SES categories to see which areas were more or less advantaged. This helped us identify the regions where students might face greater challenges accessing higher education.

Population Densities: - By overlaying demographic data on the map, we could visualize population densities. This made it easier to spot areas with a high concentration of Year 12 graduate students, highlighting potential focus areas for increasing university enrollment.

Proximity to Campuses: - One of our key goals was to see how close students from Low SES backgrounds were to Monash University campuses. Using the spatial data, we were able to create visualizations that showed which regions had better access and which might benefit from additional facilities or support.

Challenges in Geospatial Data Preparation

We faced challenges ensuring census data aligned with shapefile boundaries, especially in rural areas with sparse data. To maintain accuracy, we aggregated data from SA1 to SA3 regions. Each step required careful validation to ensure accuracy, avoiding errors in visualizations. In the end, the geospatial data allowed us to clearly visualize Victoria's socio-economic landscape, guiding Monash University in targeting support for Low SES students.

4.5 Data Analysis and Diagnostics

Our analysis examined Year 12 population trends in Victoria, focusing on socio-economic status (SES) variations. This helped identify patterns to support Monash University's Low SES enrollment goals, using R's data tools to summarize trends, categorize regions, and ensure data reliability.

Data Analysis

Summarizing Year 12 Population Trends by Region and SES

We analyzed Year 12 graduate distribution in Victoria by grouping data at the SA1 and SA3 levels, focusing on 17, 18, and 19-year-olds. Using `group_by()` in R, we segmented data by region and SES, identifying areas with high concentrations of Low SES students that might need targeted support.

Creating SES Categories with Percentile Groups

To make the socio-economic differences clearer, we divided regions into percentile groups based on their SES ranking. This categorization helped us compare areas within Victoria, ranging from the most disadvantaged to the most advantaged. We used R's `mutate()` function to create these SES categories, which grouped each region into percentile ranges from the bottom 25% to the top 25%. This step was crucial for our visualizations, as it allowed us to map socio-economic disparities across the state, highlighting regions that might struggle to meet enrollment goals.

Techniques for Grouping and Summarizing Data

We used grouping and summarizing techniques to analyze the data. The `group_by()` function in R allowed us to segment information by region, year, and SES category, and the `summarize()` function calculated averages, medians, and totals. This helped track Year 12 population changes over time across regions. Additionally, the `mutate()` function created new variables to capture trends, aiding in forecasting and assessing regional performance relative to Monash University's enrollment targets.

Diagnostics and Robustness**

Ensuring Data Reliability

To ensure data accuracy, we built several validation checks, including cross-referencing with previous census records to catch errors early. We also compared data across different geographical levels—both SA1 (local) and SA3 (regional)—to confirm that local trends aligned with broader patterns. Any discrepancies were investigated to maintain reliability, ensuring our conclusions were robust and free from isolated anomalies.

Limitations in the Analysis

While our analysis was thorough, it had limitations. We focused only on Victoria, a comparatively wealthy state, without national comparisons. Projections assumed stable government policies, but changes in funding or incentives could impact outcomes. Migration trends were also assumed to be steady, though shifts could alter projections. Additionally, our analysis depended on the accuracy of 2021 Census data, with any inaccuracies potentially affecting results, and socio-economic status can change over time, influencing stability.

Addressing the Challenges

To address limitations, we used R's flexible tools, like `mutate()` and `case_when()`, to adjust for regional patterns and handle missing data. We validated forecasts against available data and continuously updated our methods to adapt to changing socio-economic conditions. This approach ensured our analysis remained relevant, providing a clear picture of where support is needed to improve educational access for disadvantaged students. Our findings will guide Monash University in supporting Low SES students, fostering a more equitable education system.

4.6 Conceptual and Methodological Complexity

Skills and Learning

This project involved applying a variety of technical skills, which allowed us to dive deep into the data and extract valuable insights:

Data Wrangling: - We spent a lot of time cleaning and preparing data, a process known as data wrangling. The raw datasets from the Australian Bureau of Statistics (ABS) were not immediately ready for analysis—they required extensive cleaning to ensure accuracy and consistency. Using R's `tidyverse` package, we standardized formats, converted text to numbers, aligned regional identifiers, and dealt with missing or inconsistent data. - One of the biggest challenges was integrating multiple datasets that were structured differently. We had to bring together demographic data, socio-economic information, and spatial data, making sure they all lined up correctly. This meant carefully handling any discrepancies in formats and using R's powerful merging tools to stitch the data together.

Geospatial Analysis: - Mapping data was key to understanding regional differences in Victoria. We used the `sf` package in R to manage spatial data, like SA1 and SA3 boundaries, creating maps that highlighted demographic trends and SES categories. Aligning datasets was challenging, but we resolved this by transforming coordinate systems and performing spatial joins. Working with geospatial data required a mix of technical skills and creativity to ensure maps were accurate and visually clear.

Forecasting Year 12 Populations: - We forecasted the Year 12-ready population using demographic trends, applying R's `case_when()` function for flexible adjustments. Forecasting was challenging, as it required accounting for migration and socio-economic changes. We validated projections regularly by cross-checking with existing data to ensure reliability.

Data Visualization: - Presenting complex data in a clear and accessible way was essential. We used `ggplot2` in R to create a series of visualizations that combined demographic and geospatial data. This included bar charts, line graphs, and detailed maps that made it easy to interpret trends.

Challenges Faced During the Analysis

Handling complex datasets was a significant challenge as we integrated diverse demographic, socio-economic, and spatial data, each with different structures and scales. This required careful standardization of region names, formats, and coordinate systems, often involving manual adjustments to ensure everything aligned accurately. In regions with incomplete data—particularly rural areas—we aggregated information from smaller SA1 areas to broader SA3 regions to maintain a balance between specificity and coverage. Merging data from different sources also required harmonizing definitions and resolutions, using R's data manipulation tools to ensure consistency. Forecasting the Year 12-ready population added another layer of complexity, as it involved assumptions about stable demographic trends while accounting for local nuances. We continuously validated our projections, refining the models to maintain accuracy, and balanced the simplicity of the forecast with the inherent complexity of the data, ensuring that our insights were both realistic and useful.

Innovation and Novelty

We employed several innovative methods to make the analysis clearer and more effective:

Using Percentile Categories for SES Regions: - A unique aspect of our approach was categorizing Victoria's regions into SES percentile groups. This allowed us to standardize how we looked at socio-economic disparities, making it easy to compare different areas. By dividing regions into percentile groups, we could clearly see which areas were most disadvantaged and which were more advantaged.These categories made our maps more intuitive.

Combining Statistical and Geospatial Analysis: - Combining statistical analysis with geospatial visualization was a key strength of the project. Statistical analysis provided the data, while mapping gave it real-world context, revealing how socio-economic and demographic factors varied across the state.

Adaptive and Flexible Analysis: - Our methodology was flexible, using R's tools like `mutate()` and `case_when()` to adapt to new data and unexpected trends. This adaptability made our analysis

robust, ensuring insights stayed relevant and allowed us to provide up-to-date recommendations for Monash University.

4.7 Discussion of Limitations and Data Issues

While our analysis provided valuable insights into Year 12 population trends and socio-economic disparities across Victoria, it was not without challenges. This section explores some of the key limitations we encountered with the data and the potential improvements for future analysis.

Data Limitations

One of the main limitations of our analysis was the lack of detailed socio-economic data. While SEIFA rankings gave a helpful overview, they grouped regions based on factors that might have hid local differences. For example, a low SES area might still have some wealthier pockets that the data didn't show. Additionally, the SES data relied on the 2021 Census, offering a snapshot of socio-economic conditions at that time. However, these conditions can shift rapidly due to changes in local economies, policies, or unexpected events like pandemics, meaning our analysis might not reflect recent socio-economic changes that could impact Year 12 graduates. The forecasting methods also introduced limitations. Projections assumed stable demographic trends over the decade, which may not hold if migration patterns shift, government policies change, or socio-economic landscapes transform. By averaging age groups to estimate future Year 12 populations, we captured broader trends but risked missing sudden local fluctuations. Integrating multiple datasets also posed challenges, with each source having different structures and formats. Standardizing data required extensive cleaning, especially when region names didn't match perfectly. Handling missing data, often aggregating smaller regions (SA1) into larger ones (SA3), sacrificing some detail for consistency.

Future Improvements

Future work could be standardizing data formats across datasets would also enhance consistency, making regional comparisons more accurate. Incorporating dynamic, real-time data sources, like economic trends or migration patterns, would provide a more current understanding of changing conditions, making projections more adaptable and reliable. These improvements would help Monash University make better-informed, data-driven decisions to support Low SES students.

4.8 Software and Tools

Our analysis relied heavily on a suite of software, packages, and libraries designed to handle data cleaning, analysis, and visualization efficiently. Here's a breakdown of the key tools used:

Studio:

- We used RStudio, a powerful integrated development environment (IDE) for R, as our primary tool for coding, data analysis, and visualization. RStudio provided a user-friendly interface to manage the project's complex datasets, execute scripts, and visualize results seamlessly.

R Packages:

tidyverse: This collection of R packages was fundamental to our data wrangling and manipulation efforts. Packages like `dplyr`, `tibble`, and `tidyr` were used extensively to clean, format, and merge data. **sf**: For handling and visualizing geospatial data, we used the `sf` package, which enabled us to manage shapefiles and perform spatial joins. This package was crucial for creating accurate maps that displayed socio-economic trends across Victoria. **ggplot2**: A core part of the `tidyverse`, `ggplot2` was essential for creating data visualizations. We used it to generate line graphs, bar charts, and detailed maps that captured population projections and SES variations over time. **zoo**: This package was used to manage and analyze time series data, which was particularly useful for projecting Year 12-ready populations over the 2021-2030 period. **readxl**: To import and process Excel files containing demographic and socio-economic data, we relied on the `readxl` package, which facilitated seamless integration of external datasets.

These tools and packages were critical to our workflow, providing the functionality and flexibility needed to handle diverse datasets and create meaningful visualizations that guided Monash University's strategy for supporting Low SES students.

4.9 Conclusion of Methodology

Our methodology aimed to provide a clear and practical understanding of Year 12 population trends and socio-economic disparities across Victoria, all with the goal of supporting Monash University's mission to increase Low SES student enrollment as per government guidelines. By using a combination of data collection, preparation, and analysis, we built a strong foundation to explore the factors influencing educational access.

We started by gathering data from trusted sources, such as the 2021 Census, ensuring that the information was accurate and relevant. Merging and cleaning the data was a crucial step, as it allowed us to create a consistent and reliable dataset. Using R and its powerful tools, like `tidyverse` and `sf`, we were able to manage and analyze complex datasets efficiently. This mix of statistical and geospatial analysis brought the data to life, showing Monash not just the raw numbers but the broader patterns and regional differences behind them.

Our forecasting method was straightforward but effective, projecting Year 12-ready populations up to 2030 using a three-year averaging technique. Although we faced challenges, like varying data quality and the need for standardization, we addressed them by using flexible tools in R that allowed us to adapt as new information emerged.

In the end, our methodology provided Monash with valuable insights into where they can make the most impact. It highlighted areas that would benefit from targeted support, helping the university allocate resources more effectively. This approach not only painted a detailed picture of the current educational landscape but also set the stage for future efforts to make education more inclusive and accessible for all students in Victoria.

5 Results

6 Future work