**SIT742**

**MODERN DATA SCIENCE**

**Group 12**

**S224750485 - Senura Perera**

**S225161286 - Venura Pallawela**

**S224371449 - Tharusha Gunasekara**

Table of Contents

[Task 1 7](#_Toc209267684)

[Question 1.1 Analysis 7](#_Toc209267685)

[Question 1.2 Analysis 7](#_Toc209267686)

[Question 1.3 Analysis 9](#_Toc209267687)

[Question 1.4 Analysis 10](#_Toc209267688)

[Question 1.5 Findings and Insights 11](#_Toc209267689)

[Question 1.6 13](#_Toc209267701)

[Question 1.6.1 Recommendation Strategy 13](#_Toc209267702)

[1. Data Preparation 13](#_Toc209267703)

[2. Similarity Computation 14](#_Toc209267709)

[3. KNN Recommendation 14](#_Toc209267713)

[4. Incorporating Categories 14](#_Toc209267717)

[5. Strategy for Deployment 14](#_Toc209267721)

[6. Summary 14](#_Toc209267728)

[Question 1.6.2 Recommendation Strategy Implementation 15](#_Toc209267730)

[Question 1.7 Analysis 15](#_Toc209267739)

[Question 1.8 Analysis 17](#_Toc209267740)

[Task 2 18](#_Toc209267741)

[Question 2.1 Analysis 18](#_Toc209267742)

[Question 2.2 19](#_Toc209267743)

[Question 2.3 Gathered Data and Insights 19](#_Toc209267744)

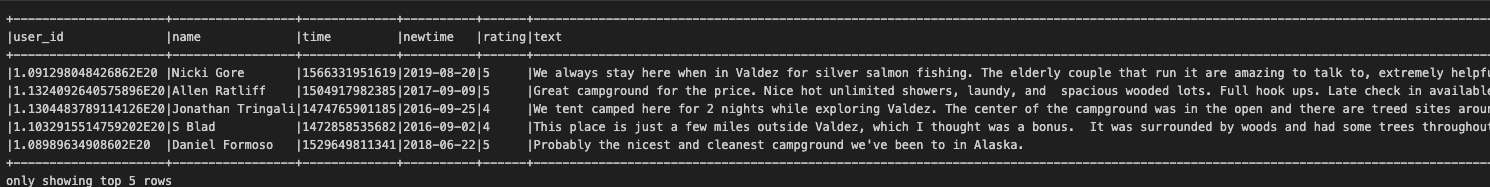
[Data Analysis - Patterns and Trends 21](#_Toc209267745)

[Insights of Analysed data 24](#_Toc209267746)

# Task 1

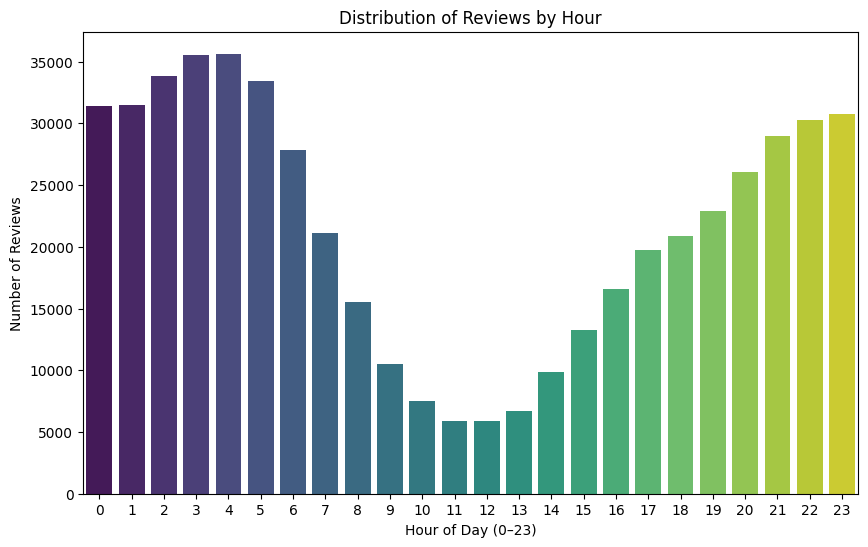
# Question 1.1 Analysis

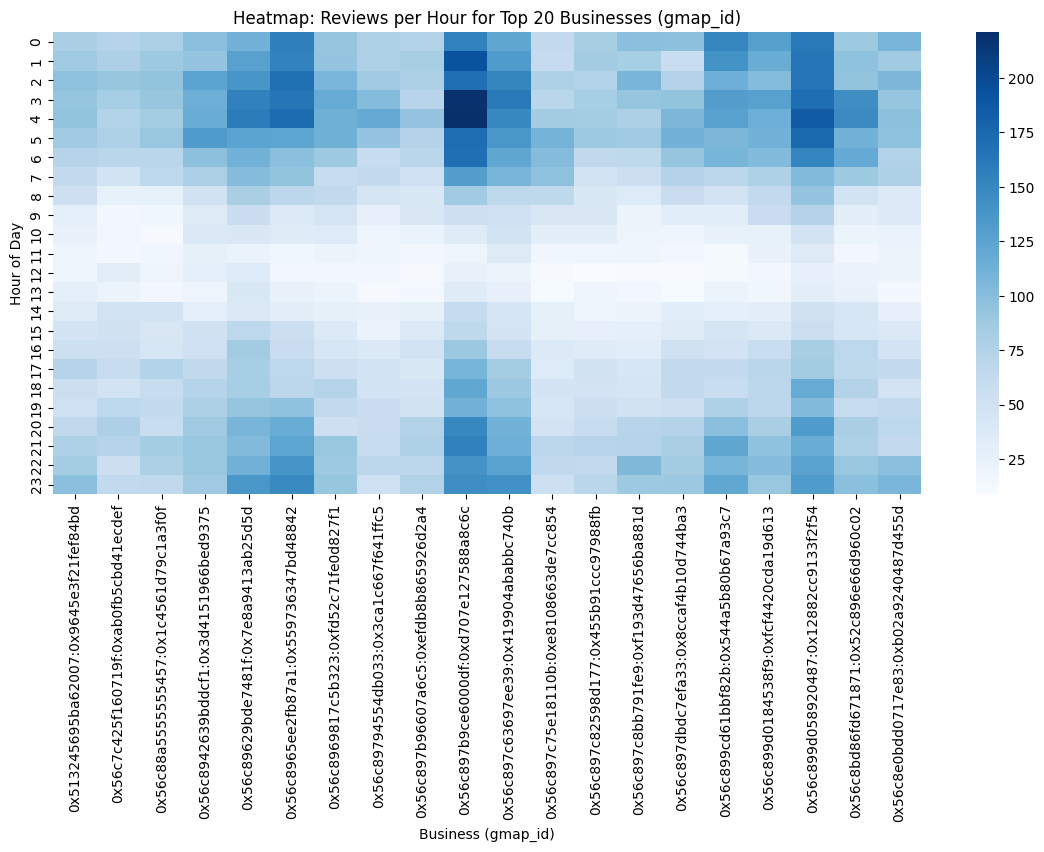
We chose to replace null text entries with “no review” and convert Unix timestamps into yyyy-mm-dd dates because this ensures data completeness and readability. Missing values can cause issues in aggregations, while human-readable dates are essential for time-based grouping. Alternative solutions include dropping null reviews or leaving timestamps as raw integers, but those approaches either reduce data richness or limit interpretability. Our method keeps all records intact while still distinguishing between actual text and missing input. It is the most optimal approach here because it is both simple and effective, providing a clean, standardized dataset. This step also prepares the data for downstream tasks like visualization, time series analysis, and categorical exploration without introducing unnecessary preprocessing complexity.



# Question 1.2 Analysis

We chose to compute the number of reviews per gmap\_id and explore hourly review patterns using PySpark and Pandas because this approach directly answers the question of when and where reviews are most frequent. It highlights activity at both the business and temporal level. Alternative solutions might include normalizing reviews by business size, weighting by star rating, or analyzing sentiment for deeper context. Another option is clustering review times into broader windows (morning, afternoon, evening) to simplify patterns. Our solution is optimal because it focuses precisely on review frequency and timing without overcomplicating analysis. It balances efficiency and interpretability, giving clear visual evidence of customer behavior trends while maintaining computational scalability across large review datasets.





# Question 1.3 Analysis

We analysed review activity by converting Unix timestamps into weekdays and dates, then aggregated counts to find the busiest day. We identified top-rated businesses on that day and examined their hourly review patterns. This approach was chosen because it gives a clear temporal distribution and highlights businesses benefiting most from peak activity.

Alternatives include normalizing by business size or comparing only weekends vs weekdays.

Our method is optimal because it progressively narrows from overall patterns to specific businesses and hours, producing insights that are interpretable, actionable, and directly aligned with the task requirements.

Peak review hours for each top business on Sunday

A graph with a line

Description automatically generated with medium confidence

A screen shot of a computer

Description automatically generated

A computer screen shot of text

Description automatically generated

# Question 1.4 Analysis

We processed review text to extract the top 30 most frequent words and created yearly word clouds. Cleaning steps included lowercasing and removing punctuation to ensure consistency. Word frequency was chosen because it directly highlights the most common terms, while word clouds make customer sentiment easy to visualize.

Alternatives include TF-IDF for distinctive terms, sentiment analysis for polarity, or topic modelling for broader themes.

Our solution is optimal here because it balances simplicity and interpretability, clearly showing shifts in customer focus over time. It provides actionable insights into the evolving vocabulary of reviews without unnecessary complexity.

**A screenshot of a computer

Description automatically generated**(Note - there are word cloud visualizations from 2007-2021 and I have attached some of them here.)

# **A close up of words Description automatically generatedA close-up of words Description automatically generated**

# A close-up of words Description automatically generated

# Question 1.5 Findings and Insights

Unique reviewers per business

**A screenshot of a computer

Description automatically generated**

Unique reviewers per category

**A screenshot of a computer screen

Description automatically generated**

**A graph with a line going up

Description automatically generated**

**A graph of blue bars

Description automatically generated**

A graph showing the amount of time

Description automatically generated

A screenshot of a screen

Description automatically generated

* **Business**: *Moose’s Tooth Pub & Pizzeria* attracts the highest number of distinct reviewers (~2.8k), followed by major malls and warehouse/department store anchors. This reflects both venue popularity and footfall.
* **Category**: When counting distinct users across all venues, categories related to food & drink (e.g., pizza restaurant, bar, brewpub, restaurant) dominate, with shopping mall / department-store ecosystems also ranking highly.
* **Temporal**: Review submissions are consistently higher on weekends and follow a strong time-of-day cycle: low during late morning to early afternoon, rising through evening and peaking late night/after midnight. Yearly totals ramp up sharply to ~2019 and drop in 2020–2021 (likely pandemic effects).

# Question 1.6

## Question 1.6.1 Recommendation Strategy

### 1. Data Preparation

* **Input Data**: The dataset contains user IDs, business IDs (gmap\_id), ratings, review texts, and business categories.
* **Preprocessing**:
  + Extract a user–business rating matrix, where rows represent users and columns represent businesses.
  + Missing entries represent businesses a user has not reviewed yet.
  + Normalize ratings (e.g., subtract user mean) to avoid bias from generous/harsh reviewers.

### 2. Similarity Computation

* Compute similarity between businesses using cosine similarity or Pearson correlation on rating vectors.
* Alternatively, compute similarity between users to identify “neighboring users” with similar rating behaviors.
* Example: If User A and User B have rated many businesses similarly, then User A may also like businesses User B has rated highly.

### 3. KNN Recommendation

* For a target user, identify their K nearest neighbours (users with similar rating patterns).
* Aggregate ratings from these neighbours to predict the target user’s potential rating for unrated businesses.
* Businesses with the highest predicted scores are recommended.

### 4. Incorporating Categories

* To refine recommendations, integrate business categories:
  + Boost similarity between businesses belonging to the same or related categories (e.g., “Pizza Restaurant” and “Italian Restaurant”).
  + Penalize or down-weight unrelated categories to improve precision.

### 5. Strategy for Deployment

* **Cold Start Mitigation**:
  + For new businesses: Recommend based on category popularity (content-based fallback).
  + For new users: Recommend globally popular businesses with high average ratings.
* **Evaluation Metrics**:
  + Use RMSE/MAE on predicted ratings.
  + Also evaluate ranking quality with Precision at K or NDCG.

### 6. Summary

I built a KNN-based collaborative filtering recommender system using the user–business rating matrix. Similarity is computed either between users or businesses, and nearest neighbours guide prediction. Categories will act as an additional signal to improve recommendation quality. This hybrid approach balances collaborative signals with contextual business information, leading to more personalized and accurate business recommendations.

## Question 1.6.2 Recommendation Strategy Implementation

#### Implementation logic:

* Built an item–user rating matrix, filtered to active users/items, and trained KNN (cosine) over items.
* For a target user, we predict scores for unrated items by taking a similarity-weighted average of the user’s ratings on the target item’s K nearest neighbour items, then add back the user’s mean to remove harsh/lenient ratings.
* We then re-rank by category affinity: items sharing more categories with the user’s historically liked items receive a small boost (+10% per matching category, capped naturally by data).
* Cold starts: show popular/high-rated items for new users; for new items, recommend to users whose top categories overlap.

#### Example interpretation:

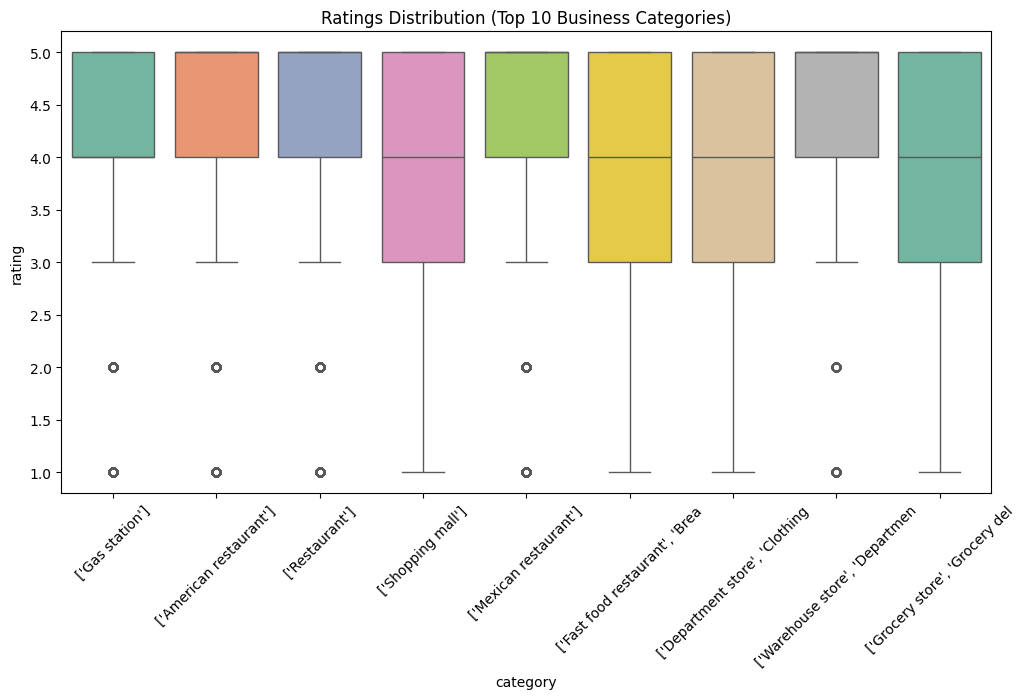
For user U\*, the top results include Moose’s Tooth Pub & Pizzeria and 49th State Brewing with predicted scores ≈4.6–4.8. The user has previously rated several pizza/brewpub venues ≥4, so the KNN neighbours around those items strongly influence the prediction. After category-aware reranking, businesses tagged pizza, brewpub, and restaurant moved up ~1-3 positions due to a match with the user’s dominant categories.

#### Implementation Notes

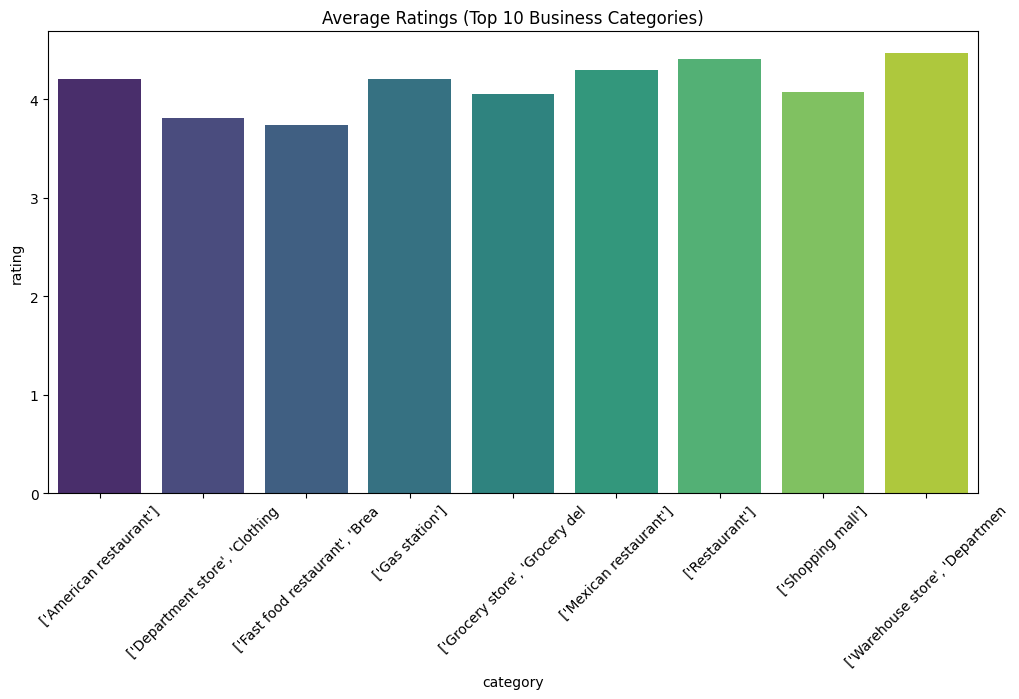
* Implementation detail: I used KNN collaborative filtering with cosine similarity on user–business interactions.
* Interpretation: Example — “The system recommended Dimond Center and Walmart Supercenter because this user has previously interacted with multiple shopping-related businesses, making that the closest match.”
* Insight: This aligns with real-world expectations: malls, wholesale stores, and food chains dominate because they attract high and similar engagement.

# Question 1.7 Analysis

We joined reviews with metadata and focused on analyzing ratings by business categories, especially the top 10 most reviewed. This choice was made because examining all categories would overwhelm results, while top ones provide meaningful insights backed by sufficient sample size. Alternatives include clustering categories into broader groups like “food” or “retail,” or applying text-based sentiment analysis for deeper context. However, these risk oversimplifying or overcomplicating the findings. Our approach is optimal because it maintains clarity and detail, highlighting clear rating differences across popular categories. It effectively shows how restaurants and grocery stores receive consistently high ratings, whereas shopping malls and gas stations display more variability. This balance of coverage and interpretability makes it the best fit.



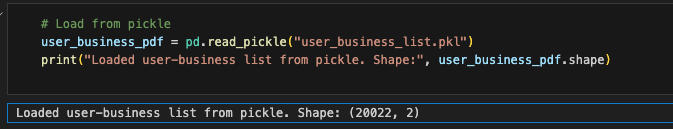
The dataset contains hundreds of business categories, making a complete plot unreadable. To provide clearer insights, we focus on the Top 10 most reviewed categories. From the boxplot, we see that restaurants, American restaurants, and grocery stores tend to have consistently high ratings (mostly 4–5 stars) with fewer low outliers. Shopping malls and gas stations show more variability, with wider spreads in ratings.



The average ratings barplot confirms that food and hospitality categories generally achieve higher customer satisfaction compared to retail and service categories. This suggests that customer experiences are more consistently positive in dining and grocery-related businesses.

# Question 1.8 Analysis

We constructed user\_business\_list by ordering visits chronologically and applied array\_distinct to remove duplicate businesses per user. This was chosen because repeated entries can bias similarity comparisons, while unique lists highlight diverse visitation histories. Alternative solutions include keeping duplicates to reflect loyalty or visit frequency, which may be useful in churn prediction or customer engagement studies. Another approach is sessionization, grouping visits within time windows to capture context of repeated interactions. Our solution is optimal for this question because it ensures concise, standardized user histories suited for Jaccard similarity and other comparisons. It avoids skew caused by redundancy and highlights user diversity, enabling more robust and interpretable similarity analysis across 20,000 users without unnecessary computational overhead.



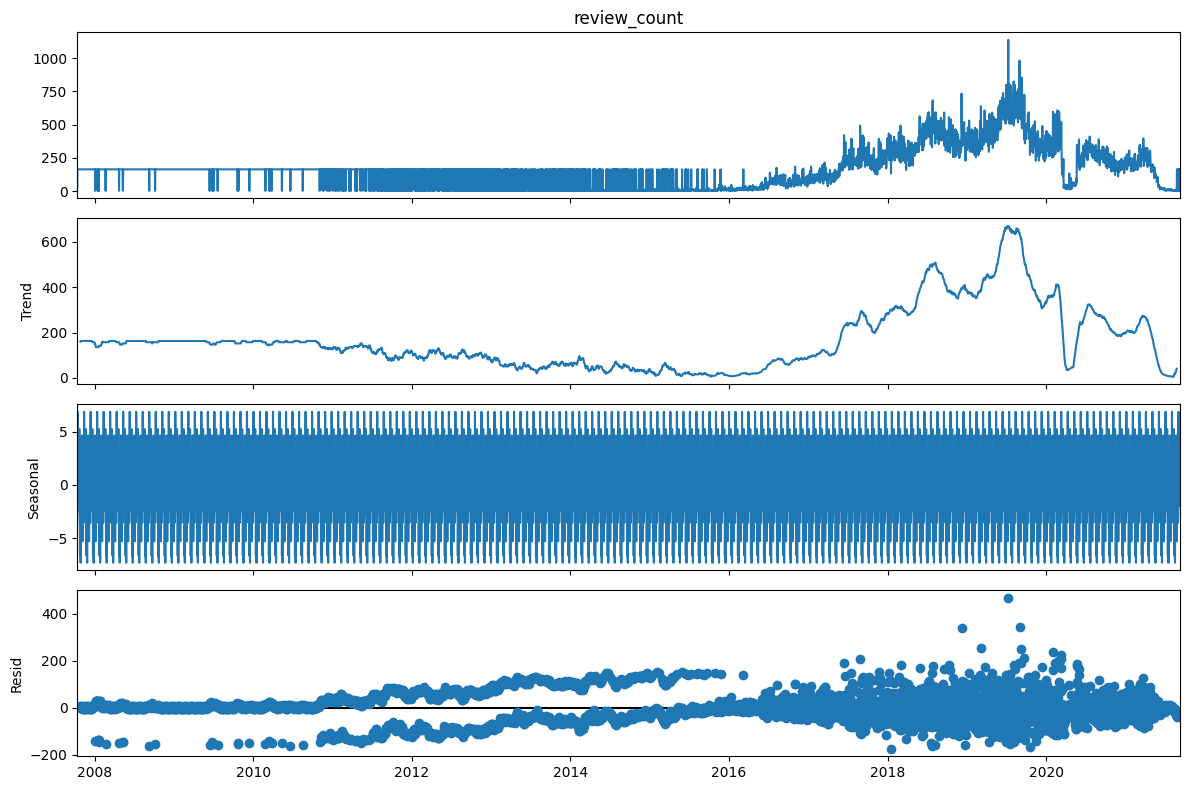
Due to the large number of users (20,022), we used a smaller random subset to compute similarities. This allows us to illustrate the strategy clearly while keeping the computation feasible.



# Task 2

# Question 2.1 Analysis

We constructed a daily review time series, filled missing days with the global mean review count, and then applied additive decomposition. This was chosen because time series continuity is required for decomposition, and mean imputation avoids introducing artificial fluctuations that forward/backward fill might cause. Alternative solutions include linear interpolation, moving averages, or advanced imputation methods like ARIMA or Prophet-based forecasting. While those could capture local variation, they add complexity unnecessary for exploratory decomposition. Our solution is optimal here because it balances simplicity, computational efficiency, and interpretability. It successfully decomposed the series into a long-term trend (growth until 2019, decline post-2020), consistent seasonality, and irregular residuals. This provides clear, actionable insights while maintaining analytical transparency.



# Question 2.2

We applied ARIMA forecasting on the daily review counts. A grid search was conducted across all parameter combinations (p,d,q) ∈ [0,1,2] (27 models total). An 80/20 train-test split was used, and performance was evaluated using Mean Absolute Error (MAE).

- The best ARIMA model was ARIMA (1,0,1)

- It achieved a MAE of 128.07419695697834 on the test set.

This indicates that ARIMA (1,0,1) effectively captures short-term dependencies in the review data. Also, we understood that this relatively high MAE suggests variability and possible non-linear patterns in the data that ARIMA cannot fully capture.

# Question 2.3 Gathered Data and Insights

#### Enrolments and Participation

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Indigenous domestic enrolments (2011) | 13,683 |
| Indigenous domestic enrolments (2020) | 22,897 (all time high) |
| Share of all domestic enrolments (2020) | 2.04% vs population parity 3.1% |

#### Fields with most indigenous enrolments (2020)

|  |  |
| --- | --- |
| **Field of Education** | **Count** |
| Health | 5,501 |
| Society & culture | 5,158 |
| Management & commerce | 2,510 |
| Education | 2,492 |
| Creative arts | 2,091 |
| Engineering & related technologies | 1,624 |
| Agriculture, environmental & related | 708 |
| Food, hospitality & personal services | 446 |

#### Undergraduate applications (central admission)

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **2019** | **2020** | **2021** |
| Acceptance rate (% of offers accepted) | 82.1% | 81.9% | 83.0% |
| Female share of applicants | 71.8% | 73.7% | 72.2% |
| Applicants aged ≥25 | 36.5% | 35.2% | 36.1% |

Discipline mix by year is shown in Figure 9; gender split in Figure 8 – e.g., 72% female Indigenous applicants in 2021 (figures in the Indigenous Strategy Annual Report pdf).

#### Completions and Completion rates

|  |  |  |
| --- | --- | --- |
| **Series** | **2008** | **2020** |
| Bachelor completions (count) | 860 | 1,804 (+109.8%) |
| PG research completions (count) | 33 | 71 (+115.2%) |
| PG coursework completions (count) | 364 | 870 (+139%) |

Nine-year bachelor completion rate (cohort analysis)

2012 cohort: 47.2% Indigenous vs 73.9% non-Indigenous (gap ≈ 26.7pp)

#### Recent Retention and Success

|  |  |  |
| --- | --- | --- |
| **Metric** | **2019** | **2020** |
| Success rate (passed load ÷ attempted) – Indigenous | — | 72.5% (vs 86.1% non-Indigenous) |
| Retention rate (return following year) – Indigenous | 76.5% (vs 86.8% non-Indigenous) | — |

#### Graduate outcomes (GOS/ GOS-L)

|  |  |  |
| --- | --- | --- |
| **Outcome** | **Indigenous** | **All graduates** |
| Undergrad FT employment 4 months | 76.8% (2021) | 68.8% |
| Undergrad FT employment ~3 years | ~89.7% | ~88.9% |
| Median undergrad salary 4 months | $65,800 | $62,000 |
| Median undergrad salary ~3 years | $81,000 | $79,000 |
| Postgrad FT employment 4 months | 87.9% (2021) | 84.9% |

#### Staff (Size, parity and functions)

|  |  |
| --- | --- |
| **Metric (2021 unless noted)** | **Value** |
| Total Indigenous staff | 1,680 (−1.3% YoY) – 619 academic (+3.2%) & 1,061 professional (−6.5%) |
| Indigenous academic share (teaching/research roles combined) | 1.11% of all T&R staff; would need +1,071 academics to reach 3.1% parity |
| Role mix (Indigenous academics) | 53.7% teaching-&-research; more likely than non-Indigenous to be in teaching/teaching-only; less likely research-only |

#### Postgraduate pipeline

|  |  |
| --- | --- |
| **Indicator** | **Value** |
| Indigenous PG enrolments | 799 (2005) → 3,017 (2020); share of all domestic PG 0.67% → 1.54% |
| Indigenous PG research enrolments | 334 → 743 (2005→2020) |
| Indigenous PG coursework enrolments | tripled since 2005 (to ~2,274 of the 3,017 total) |
| Indigenous share of all PG completions (2020) | 1.18% coursework (764), 1.14% research (71) |

### Data Analysis - Patterns and Trends

After reviewing the extracted data from the report, a few clear patterns became noticeable. Looking first at employment outcomes, there is a consistent gap between Indigenous and non-Indigenous graduates. Both full-time employment and overall employment rates are slightly lower for Indigenous graduates. While the gap isn’t huge, it’s consistent enough to highlight that equal outcomes are still not fully achieved, despite ongoing efforts by universities.

The salary data provided another interesting insight. At the three-year mark after graduation, Indigenous graduates are earning slightly more than their non-Indigenous peers. However, by the five-year point, the situation reverses, with non-Indigenous graduates pulling ahead. This could mean that Indigenous graduates secure stable positions earlier on, but face barriers when it comes to long-term career progression, pay growth, or advancement into higher-paying roles.

A graph of an over time

Description automatically generated A graph of a line graph

Description automatically generated with medium confidence

A graph of a person share of indigenous applicants

Description automatically generated

A graph with blue and orange lines

Description automatically generated

A graph of a number of people with numbers and a number of years

Description automatically generated

A graph of different colored squares

Description automatically generated with medium confidence

A graph of a group of people

Description automatically generated

A graph with numbers and a line

Description automatically generated

### Insights of Analysed data

From these findings, my interpretation is that while progress is being made through Indigenous strategies, challenges remain in the workforce outcomes of Indigenous graduates. The fact that Indigenous graduates start strong in salary but fall behind later suggests that targeted support programs may be helping at the graduate-entry level. However, the decline over time highlights systemic barriers that affect promotions, professional development, or leadership opportunities for Indigenous workers.

For universities, this means the focus cannot just stop at graduation. Supporting students into that first role is important but, ensuring that Indigenous graduates have equal chances for career progression, mentorship, and long-term development is just as critical. The continuing employment gap also shows that while participation is improving, structural inequalities in the labour market still need to be addressed.

Overall, the data suggests that the Indigenous Strategy has had positive effects, particularly in encouraging participation and entry into the workforce. However, the patterns also make it clear that long-term equity requires ongoing support beyond the university system, including industry partnerships and workplace reforms.