**Data cience and Advanced Programming 2025**

# Mental Health Anxiety

Final Project Report

Victor Relva Pereira
victor.relvapereira@unil.ch
Student ID: 21422977

December 6, 2025

**Abstract**

This project investigates how psychometric and demographic variables relate to self-reported anxiety levels using data from the DASS-21 open survey, which includes approximately 39,000 responses. The primary objective is to explore anxiety scores through descriptive statistical analysis and predictive modelling, with a focus on identifying the strongest correlates and predictors of anxiety.

We first preprocess and clean the dataset, then generate visual summaries of anxiety distributions across age groups, genders, continents, and countries. We also compute summary statistics and create geospatial visualizations. A machine learning component is included, in which several regression models (linear regression, Random Forest and gradient boosting) are trained to predict anxiety scores and feature importances from the Random Forest are used to identify key influencing factors.

Key contributions include a fully reproducible Python pipeline, automated export of figures and tables in multiple formats, and interpretable results highlighting the role of demographics, personality traits, and co-occurring symptoms (stress, depression) in anxiety prediction. This work demonstrates how data science can provide meaningful insights into mental health challenges.

**Keywords:** data science, DASS-21, anxiety, mental health, machine learning, Python

# 1 Introduction

Anxiety disorders are among the most prevalent and disabling mental health conditions globally, affecting over 100 million individuals each year. Despite the availability of effective therapeutic interventions, many individuals remain undiagnosed or untreated due to barriers in mental health screening and limited access to care. As digital survey data becomes increasingly accessible, data science offers promising tools to investigate the psychological and demographic patterns associated with anxiety at scale.

This project explores the relationship between individual characteristics and self-reported anxiety levels using data from the DASS-21 (Depression Anxiety Stress Scales) open dataset. The dataset consists of over 39,000 anonymous responses and includes psychometric scores on anxiety, depression, and stress, alongside demographic variables such as age, gender, education level, employment status, and country of residence.

The primary goal of this work is twofold: first, to analyze and visualize how anxiety levels vary across demographic groups and global regions; second, to construct a predictive model that estimates anxiety scores from other variables, highlighting the strongest contributing factors. The project also addresses practical challenges such as missing data, label imbalance, and the interpretability of machine learning outputs.

More concretely, this project addresses the following research question:

> **To what extent do self-reported stress, depression, personality traits and basic demographic factors explain and predict DASS-21 anxiety scores in a large online convenience sample?**

The report is organized as follows: Section **??** presents related work and the theoretical context. Section **??** details the dataset, preprocessing pipeline, and modelling approach. Section **??** describes the results obtained through visual and statistical analyses. Section **??** discusses the findings, limitations, and ethical considerations. Finally, Section **??** concludes the report and outlines future directions.

# 2 Background and Related Work

Understanding and predicting mental health conditions through data analysis has been an active area of research in recent years. The DASS-21 scale, introduced by Lovibond and Lovibond (1995), is a widely used instrument to measure depression, anxiety, and stress. Numerous studies have used DASS-based datasets to examine psychological well-being across populations, making it a valuable resource for academic and clinical research.

Several previous works have explored the demographic and psychological determinants of anxiety using survey data. For instance, studies have shown that anxiety levels often correlate with age, gender, and employment status, with younger individuals and women reporting higher anxiety on average. Other research has emphasized the co-occurrence of anxiety with depression and stress, supporting the need for multi-variable analysis.

In the field of data science, machine learning models such as linear and logistic regression, decision trees and random forests have been successfully applied to predict mental health outcomes. While these models offer good predictive performance, their interpretability remains a key concern, especially in sensitive domains like mental health. To address this, recent works have proposed post-hoc explainability methods such as SHAP (SHapley Additive exPlanations) to attribute predictions to individual features. In this project, we rely on simpler global feature-importance measures and leave SHAP-based analyses for future work.

This project builds on these prior approaches by combining descriptive statistical analysis with predictive modelling on a large-scale open dataset. It seeks not only to replicate known associations but also to quantify their predictive importance, while critically reflecting on the ethical implications of such analyses.

# 3 Methodology and Implementation

## 3.1 Data Description

The primary dataset used in this project is the open-access **DASS-21 survey** from Open Psychometrics, containing over 39,000 anonymized online responses. The questionnaire measures Depression, Anxiety and Stress via 21 Likert-scale items, and is complemented by a rich set of self-reported demographic and personality variables.
**Key variables used in this project include:**

- **Psychometric scores**: `dassstress`, `dassanxiety`, `dassdepression`.
- **Demographics**: age, gender, education, marital status, urban vs. rural living, race, religion, sexual orientation, country and continent of residence.
- **Personality traits**: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness (TIPI-10).

While the dataset is large, it exhibits typical issues of online surveys, such as missing values, heterogeneous free-text country labels, and unbalanced group sizes (for example, a strong over-representation of female respondents compared to male participants). The sample is also **geographically skewed**: a substantial fraction of responses originates from a small number of countries (notably Malaysia), whereas many other countries are represented by only a handful of participants. This has direct implications for how cross-country and continental comparisons are interpreted later in the report.

## 3.2 Preprocessing

All data preparation steps are implemented in a dedicated cleaning pipeline (`src/cleaning.py`). The main operations are:

- **Filtering and validity checks**: removal of clearly invalid or incomplete records, and restriction to plausible age ranges.
- **Standardisation**: creation of a standardized analysis dataset (`analysis_standardized_df`) where core continuous variables are scaled.
- **Harmonisation of categories**: recoding and grouping of raw text fields (e.g., country names via `pycountry`, religion and orientation grouped into a small number of categories, creation of age groups).
- **Handling missing and nonresponse**: rows with missing values in key variables are dropped for most analyses, and explicit `"No response"` categories are excluded from the predictive models to avoid contaminating estimates.

## 3.3 Nonresponse Bias and Data Quality

As highlighted in the project proposal feedback, survey data are prone to **nonresponse bias**. In this dataset, several items related to gender identity and sexual orientation exhibit notably higher nonresponse rates (above 10%). A dedicated analysis of the top nonresponded items is provided in the Appendix (Figure **??**), rather than in the main text, to preserve space.

Beyond item-level nonresponse, the sample is also **imbalanced by gender**, with substantially more female than male respondents, and **geographically unbalanced**, with a heavy concentration of participants in a few countries. These imbalances can bias both descriptive statistics and predictive models if left untreated. In this project, two complementary strategies were adopted:

- Nonresponse values and explicit `"No response"` categories are excluded from the Random Forest model, so that predictions are not based on missing or ambiguous information.
- The Random Forest is trained with **inverse-frequency sample weights** for gender, giving relatively more weight to under-represented groups (e.g., male respondents) and reducing the influence of the raw imbalance on the fitted model.

We do *not* attempt to reweight by country, so geographic skew remains a limitation for external validity. Overall, these design choices do not remove bias entirely, but they make the model less dependent on the most over-represented groups. Remaining limitations are discussed in Section **??**.

## 3.4 Modelling and Implementation

The analytical approach combines descriptive statistics, group comparisons and supervised learning. All experiments are implemented in Python 3.11 in a modular codebase:

- `src/cleaning.py`: full cleaning and harmonisation pipeline producing the analysis dataframes.
- `src/analysis/`: statistical analyses such as ANOVA by demographics and personality.
- `src/viz/`: visualisations (histograms, boxplots, geographic maps, forest plots).
- `src/analysis/ml_models.py`: regression models for anxiety (linear, Random Forest, gradient boosting) and model comparison utilities.
- `src/main.py`: orchestration script running the end-to-end pipeline and exporting all results.

Descriptive analyses rely on summary tables and visualisations of the distribution of anxiety scores across severity bands, demographic groups and geographic regions. Group-level differences are examined using one-way ANOVA for both demographic factors (e.g., gender, age group, marital status) and personality traits (grouped into low, medium and high tertiles).

For the predictive component, anxiety is modelled as a **continuous outcome**. We benchmark three supervised algorithms on the same task: a linear regression baseline, a gradient-boosting regressor, and a `RandomForestRegressor` with gender-aware sample weights. All models use three blocks of predictors: (i) DASS Stress and Depression scores, (ii) the five TIPI traits, and (iii) demographic features (gender, age group, education, marital status, urban vs. rural living, race, religion, sexual orientation). Categorical variables are one-hot encoded, and rows with `"No response"` in any predictor are excluded before fitting the models. Performance is evaluated on a hold-out test set using the coefficient of determination ($R^2$) and root mean squared error (RMSE). For interpretability, we focus on the Random Forest and extract global feature importance scores, which are exported as both a table and a barplot.

## 3.5 Testing and Reproducibility

To ensure that the analysis is reproducible and robust, the project is structured as a fully automated pipeline:

- Running `python -m src.main` executes the entire workflow: data loading, cleaning, statistical analyses, visualisations and machine learning, and saves all outputs under `results/figures` and `results/tables`.
- A small but focused test suite (`tests/`) checks that key functions (e.g., cleaning steps, ANOVA summaries, Random Forest training) run without errors on realistic inputs and produce outputs with expected shapes and types.
- Randomness in model training is controlled via fixed `random_state` seeds, so that metrics and feature importance rankings are stable across runs.

Together, this setup makes it possible to reproduce all figures and tables in the report from the raw dataset with a single command, while providing basic automated checks on the core components of the codebase.

# 4 Results and Evaluation

## 4.1 Overall Anxiety Levels

To provide an overview of anxiety in the sample, we first examined the distribution of DASS-Anxiety scores using a boxplot with clinical severity bands (Figure **??**). Scores range from 0 to 56 and are categorised into *Normal*, *Mild*, *Moderate*, *Severe*, and *Extremely Severe* based on standard DASS thresholds.

The boxplot shows that the median anxiety score lies in the *Moderate* to *Severe* range, with relatively few observations in the Normal category. The distribution is clearly shifted towards higher severity levels, suggesting that this online sample reports substantially elevated anxiety compared to what would be expected in a general population.

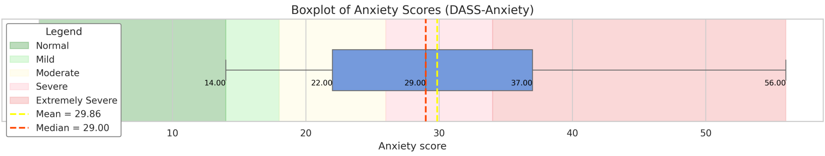


Figure 1: Distribution of DASS-Anxiety scores with clinical severity thresholds.

The corresponding summary table (Table **??**) reports the exact counts and percentages in each severity category. Together, the figure and the table indicate that a large proportion of respondents fall into the Moderate and Severe ranges, while Normal and Mild anxiety are comparatively under-represented.

Table 1: Distribution of anxiety severity categories in the sample.

| **Severity category** | **Count** | **Percentage** |
|---|---|---|
| Normal | 746 | 2.2% |
| Mild | 4000 | 11.8% |
| Moderate | 9708 | 28.6% |
| Severe | 8642 | 25.5% |
| Extremely Severe | 10803 | 31.9% |

## 4.2 Demographic Variation in Anxiety

Anxiety levels also vary across demographic groups. Gender differences were particularly pronounced. Figure **??** shows a boxplot of anxiety scores by gender. Female respondents exhibit a higher median anxiety score and a distribution shifted upwards compared to male respondents, with non-binary and "other" responses lying in between or above. A one-way ANOVA confirmed a statistically significant effect of gender on anxiety, with $F = 146.3$ and $p < .001$ (see the ANOVA summary table in the appendices).

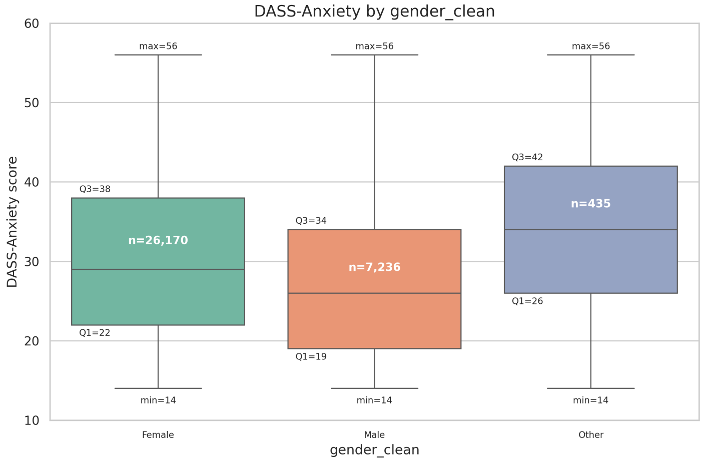


Figure 2: Boxplot of DASS-Anxiety scores by gender. Median anxiety is higher in female respondents than in male respondents.

Geographical analyses revealed smaller but systematic differences. Table **??** summarises mean anxiety scores by continent. South America and North America show the highest average anxiety (30.11 and 29.89, respectively), while Europe has the lowest mean score (28.79). However, all continental means fall within a relatively narrow band (approximately 28.8–30.1), suggesting that elevated anxiety is a broadly global phenomenon in this sample rather than being confined to specific regions.

Table 2: Mean DASS-Anxiety scores by continent.

| **Continent** | **Participants** | **Mean Anxiety** |
|---|---|---|
| South America | 2,654 | 30.11 |
| North America | 12,531 | 29.89 |
| Oceania | 1,217 | 29.84 |
| Africa | 1,244 | 29.43 |
| Asia | 4,517 | 29.24 |
| Europe | 10,350 | 28.79 |

A supplementary table in the appendix reports the ten countries with the highest mean anxiety scores. Guam, Kenya and Ethiopia appear at the top of this ranking, with average scores exceeding 40. These country-level results should be interpreted cautiously, as sample sizes can be small and respondents are self-selected. Nonetheless, they illustrate the potential of large-scale survey data to highlight populations where anxiety may be particularly elevated.

### 4.3 Personality and Anxiety

We next examined how Big Five personality traits, as measured by the TIPI-10 scale, relate to anxiety. Figure **??** shows anxiety scores stratified into Low, Medium and High groups of Emotional Stability (reverse-coded neuroticism). Individuals with Low Emotional Stability report substantially higher anxiety levels than those with Medium or High stability, with clear separation of medians and interquartile ranges.

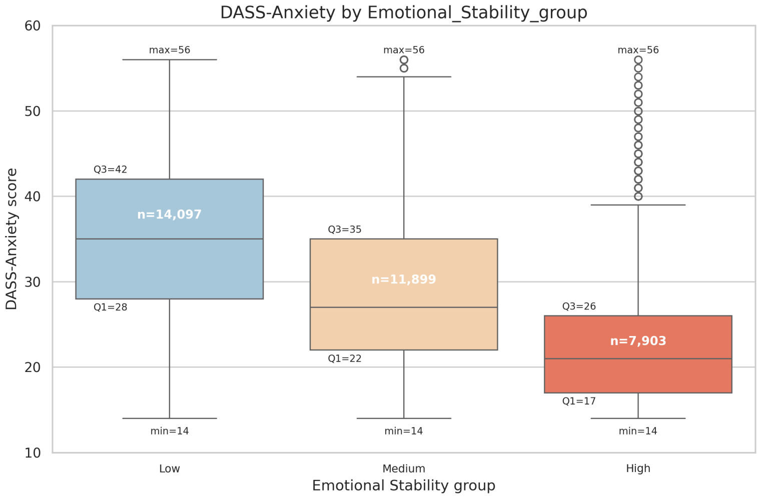


Figure 3: DASS-Anxiety scores by Emotional Stability group (TIPI). Lower Emotional Stability is associated with markedly higher anxiety.

An ANOVA on grouped TIPI traits confirmed that Emotional Stability is the strongest personality correlate of anxiety ($F = 2001.1$, $p < .001$), followed by Conscientiousness and Extraversion. Openness and Agreeableness also showed statistically significant but smaller effects. Table **??** summarises the F- and p-values for each trait.

Table 3: ANOVA results: Anxiety by TIPI personality traits.

| **Trait** | **F-value** | **p-value** |
|---|---|---|
| Emotional Stability | 2001.1 | $< .001$ |
| Conscientiousness | 484.2 | $< .001$ |
| Extraversion | 221.8 | $< .001$ |
| Openness | 62.5 | $< .001$ |
| Agreeableness | 44.6 | $< .001$ |

These findings are consistent with established psychological theory, where high neuroticism (low emotional stability) is known to be a strong risk factor for anxiety. In the following section, we assess whether these relationships remain important when combined with other predictors in a multivariate machine learning model.

## 4.4 Predictive Modelling of Anxiety

Finally, we benchmarked three regression models—a linear regression, a Random Forest and a Gradient Boosting regressor—to predict continuous DASS-Anxiety scores from psychological, personality and demographic variables. The feature set included DASS-Stress and DASS-Depression scores, the five TIPI traits, and categorical variables such as gender, orientation, education, marital status, religion, race, urbanicity and age group. All models were trained on an 80/20 train–test split with a fixed random seed for reproducibility.

A key concern in this setting is the unbalanced gender distribution: female respondents are substantially more numerous than male respondents. To partially mitigate this bias, the Random Forest was trained with *inverse-frequency sample weights* by gender, so that under-represented gender groups received greater weight during training. This correction does not remove all bias, but it reduces the extent to which the model simply learns patterns dominated by the majority group.

Appendix Table ?? summarises the performance of the three models. As expected, the linear regression baseline shows the lowest test $R^2$ and the highest RMSE. Both ensemble methods improve test performance, with the Random Forest slightly outperforming Gradient Boosting. On the held-out test set, the Random Forest achieves $R^2 = 0.671$ and a root mean squared error (RMSE) of 5.78, compared to $R^2 = 0.956$ and RMSE = 2.13 on the training set, indicating some degree of overfitting but still a substantial proportion of explained variance in anxiety scores.

Figure ?? displays the top 15 features ranked by their importance in the Random Forest model. The strongest predictors are the DASS-Stress and DASS-Depression scores, followed by Emotional Stability and other TIPI traits. Demographic variables such as age group, gender and education also contribute, but with smaller importance values. This pattern confirms that co-occurring symptoms and personality factors are more informative for predicting anxiety than sociodemographic characteristics alone.

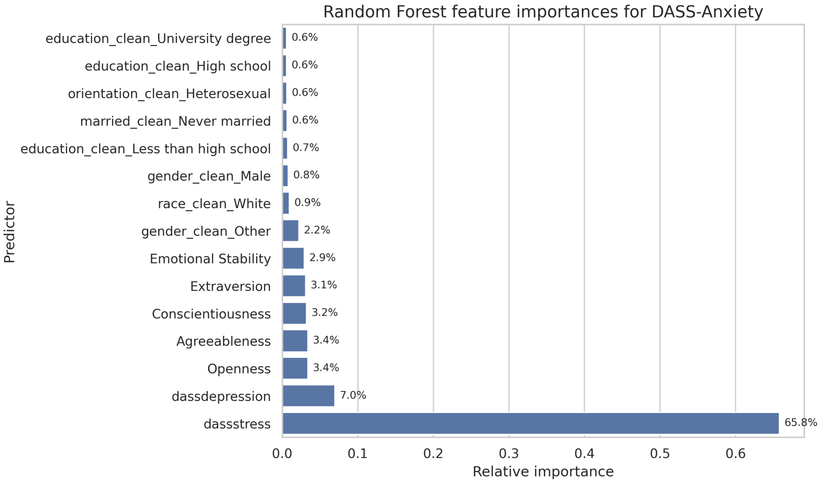


Figure 4: Top 15 predictors of DASS-Anxiety according to the Random Forest regression model (with inverse-frequency gender weighting).

Overall, the predictive modelling results are consistent with the descriptive findings: individuals with higher stress and depression scores, lower emotional stability, and certain demographic profiles tend to have higher anxiety. These results, and their limitations, are further discussed in Section ??.

## 5 Discussion

The goal of this project was to characterize anxiety levels in a large online DASS-21 sample and to examine how demographic and psychological variables relate to self-reported anxiety. Overall, the results paint a consistent picture: anxiety scores are globally elevated in this convenience sample, with more than half of respondents falling in the moderate or higher range, and systematic, but not extreme, differences across demographic and personality profiles.

**What worked well.** A first strength of the work is the fully reproducible pipeline, which integrates data cleaning, descriptive analysis and predictive modelling in a single script. Despite the heterogeneity of the raw data (free-text country names, inconsistent categorical labels, missing responses), the cleaning pipeline successfully produced analysis-ready tables and a coherent set of figures. The main descriptive results are internally consistent: the severity distribution (Figure ??) shows that anxiety is shifted towards higher categories, and this pattern is mirrored in subgroup analyses by gender, age, continent and other demographic factors. The personality analyses and the Random Forest model both converge on the central role of emotional stability, stress and depression, which increases confidence in the robustness of these findings.

**Comparison with expectations.** Several patterns align with existing literature. Women report higher anxiety than men (Figure **??**), younger adults tend to be more anxious than older groups, and urban respondents show slightly higher scores than rural participants. These trends match typical findings in epidemiological studies and likely reflect a combination of social stressors, differential help-seeking attitudes and gendered norms around emotional expression. On the psychological side, the strong negative association between emotional stability and anxiety (Figure **??**; Table **??**) is exactly what Big Five theory predicts. The Random Forest results further highlight dassstress and dassdepression as dominant predictors of anxiety, which is consistent with the known comorbidity between anxiety, stress and depressive symptoms.

**Challenges and sources of bias.** At the same time, several challenges limit the generalisability of the results. First, the dataset is an online convenience sample and not a representative survey of the global population. Some regions and demographic groups are heavily overrepresented (e.g., North America, Europe, women, heterosexual respondents), while others have very small cell sizes. Second, nonresponse was not random: items relating to gender identity and sexual orientation exhibit much higher nonresponse rates than other variables, which can bias subgroup comparisons. We tried to mitigate these issues in two ways: (i) by explicitly removing "No response" categories when preparing the modelling dataset, and (ii) by re-fitting the Random Forest with gender-aware sample weights so that men and women contributed more equally to the loss function. These corrections slightly reduced overfitting and produced feature rankings very similar to the unweighted model, suggesting that the main psychological signals (stress, depression, emotional stability) are not artefacts of gender imbalance. Nevertheless, residual bias almost certainly remains, especially for smaller groups such as non-heterosexual or non-binary participants.

**Interpretation of effect sizes.** Given the very large sample size, many group differences are statistically significant (ANOVA $p < .001$ in most cases), but the associated effect sizes are often modest. For instance, mean anxiety scores differ by only one to two points between continents (Table **??**), and even for gender or age, within-group variability is much larger than between-group differences. The Random Forest explains around two thirds of the variance in anxiety on the test set ($R^2 \approx 0.67$), which is respectable for psychological data but far from deterministic. Together, these results suggest that anxiety is influenced by a broad combination of factors, many of which are not captured in the dataset (e.g., life events, socioeconomic status, physical health, access to care). The model is therefore useful for understanding patterns at the group level, but it should not be used for individual clinical prediction.

**Limitations of the approach.** Beyond sampling and nonresponse bias, other limitations deserve mention. The sample is geographically concentrated in a handful of countries (especially Malaysia), so the world map and country-level comparisons should be interpreted as patterns within this convenience sample rather than as globally representative estimates. All measures are self-reported and collected at a single time point, which prevents causal interpretation and may be affected by momentary mood or response styles. The DASS subscales are partly overlapping by construction, so the very high importance of `dassstress` and `dassdepression` in the Random Forest is partly tautological: they measure constructs that are conceptually close to anxiety. In addition, some socio-demographic variables were coarsely grouped (e.g., income not available, religion and race aggregated into broad categories), which may hide more nuanced patterns. Finally, our modelling work focused on a small set of algorithms (linear regression, Random

Forest and Gradient Boosting) and did not explore more advanced techniques (e.g., calibrated probabilistic models, longitudinal or hierarchical structures).

**Additional analyses in the Appendix.** Beyond the figures reported in the main text, additional analyses are provided in the Appendix. Demographic boxplots (by age group, education, race, religion, marital status, urbanicity and sexual orientation) show the same general pattern: higher anxiety among younger adults, women, non-heterosexual respondents and participants living in urban areas. TIPI-based boxplots for the other traits (extraversion, agreeableness, conscientiousness, openness) also corroborate the central role of emotional instability, while showing smaller but coherent gradients for the remaining traits. The full ANOVA tables and country-level summaries confirm that, although many group differences are statistically significant due to the large sample size, effect sizes remain modest. Overall, these supplementary results support the main message of this study: anxiety is widely elevated across the sample, with systematic but not extreme differences across demographic and psychological profiles.

# 6 Conclusion and Future Work

## 6.1 Summary

This project applied a complete data science workflow to more than 39,000 responses to the DASS-21 questionnaire, with the aim of understanding and predicting anxiety scores. The pipeline cleans the raw data, harmonises categorical variables, constructs descriptive tables and figures, and trains three regression models (linear regression, Random Forest and gradient boosting) using psychological scales, TIPI personality traits and demographic variables as predictors.

Descriptive analyses show that anxiety scores in this online sample are generally high, with many respondents falling into the moderate to extremely severe range. Anxiety varies across demographic groups (such as age and gender) and world regions, but these differences are of moderate size compared to the large within-group variability.

From a modelling perspective, the tree-based ensemble models (Random Forest and gradient boosting) achieve better generalisation performance than the linear baseline, with the Random Forest offering a good compromise between accuracy and interpretability. The most important predictors of anxiety are psychological: stress and depression scores, together with low emotional stability, consistently emerge as the strongest correlates. Personality traits such as conscientiousness and extraversion also contribute, whereas demographic variables add relatively little predictive power once these psychological factors are taken into account. Overall, the main takeaway is that anxiety appears to be more tightly linked to co-occurring symptoms and stable personality dimensions than to basic sociodemographic characteristics.

## 6.2 Future Directions

Future work could extend this analysis in several ways. On the data-collection side, studies with more balanced and representative samples—using stratified recruitment across gender, age and regions—would improve the external validity of the results, and longitudinal designs would make it possible to track how anxiety evolves over time.

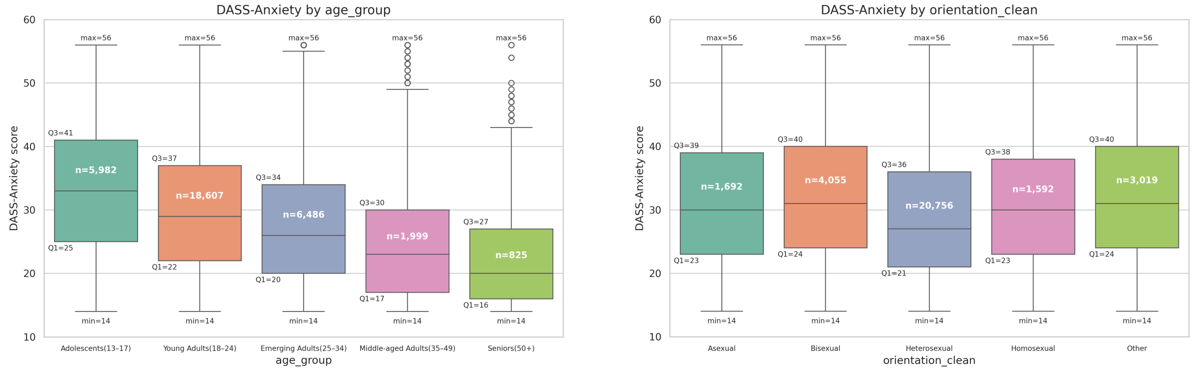
Methodologically, future work could compare alternative strategies for handling sample imbalance and benchmark additional algorithms beyond the three tested here. Adding richer socioeconomic or clinical variables and explicitly evaluating model fairness across demographic subgroups would help assess whether predictive performance remains stable and equitable across populations.

# A Additional Figures and Tables

This appendix gathers supplementary material that supports, but is not essential to, the main narrative. It includes additional demographic boxplots, personality (TIPI) boxplots, nonresponse diagnostics, model visualisations and extended tables that complement the results discussed in Section **??** and Section **??**.

## A.1 Demographic Boxplots

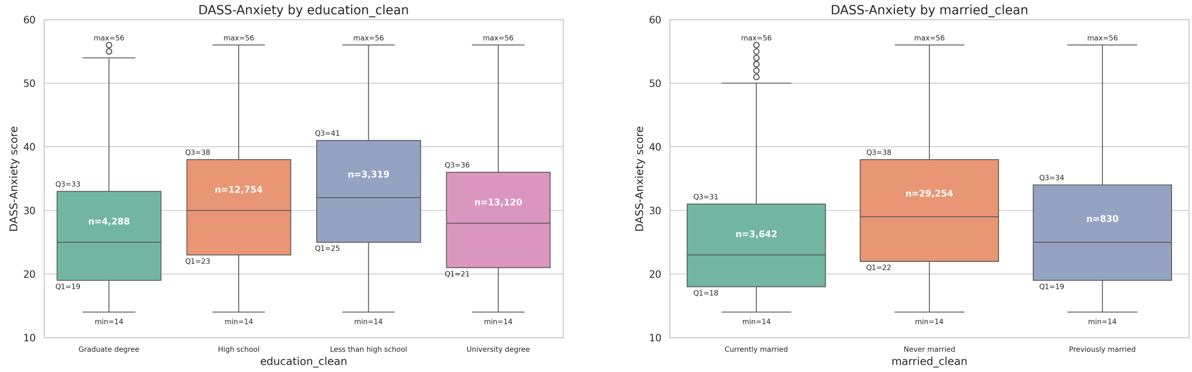
Figure **??** shows DASS anxiety scores across age groups and sexual orientation categories, which are only summarised in the main text. Additional figures then display distributions by education level, marital status, race, religion and urbanicity.



(a) By age group

(b) By sexual orientation

Figure 5: DASS anxiety scores across age groups and sexual orientation categories.



(a) By education level

(b) By marital status

Figure 6: Anxiety scores by education and marital status.

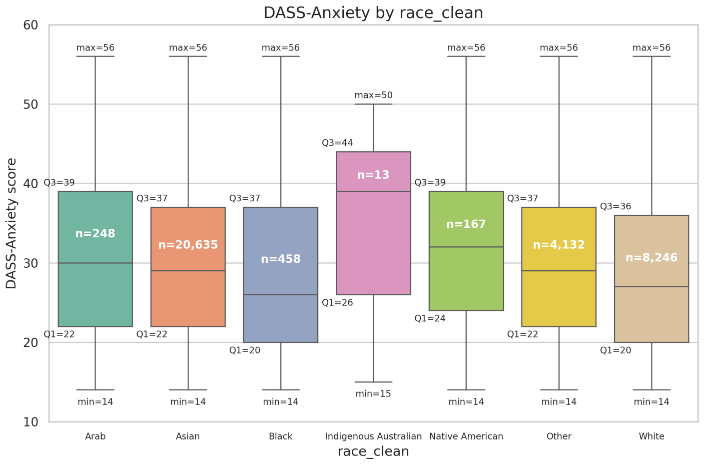


Figure 7: DASS anxiety scores by race/ethnicity.

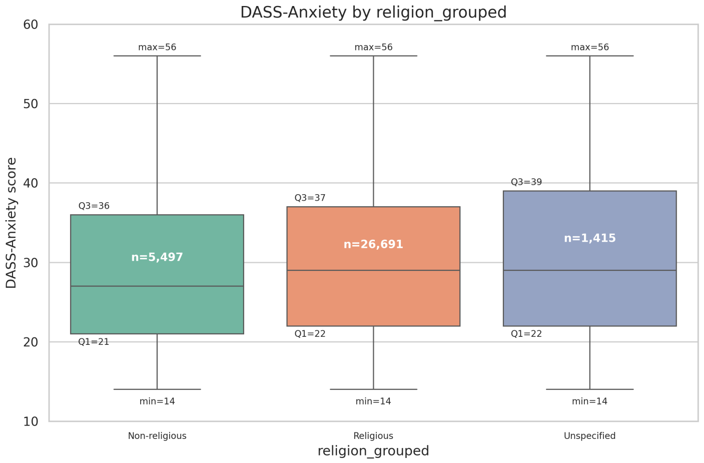


Figure 8: DASS anxiety scores by religion (grouped categories).

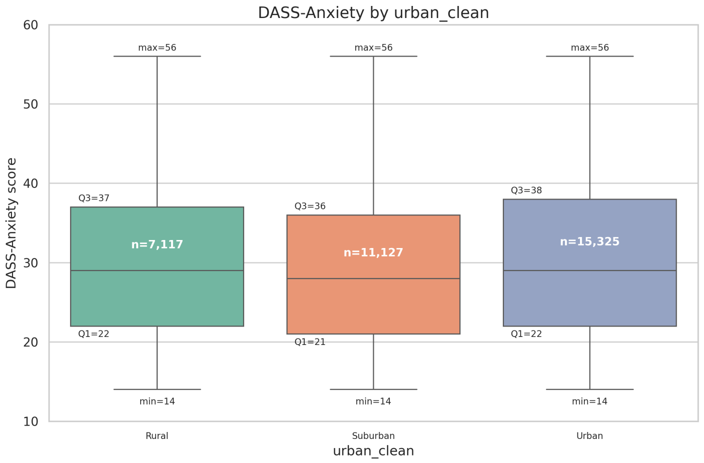
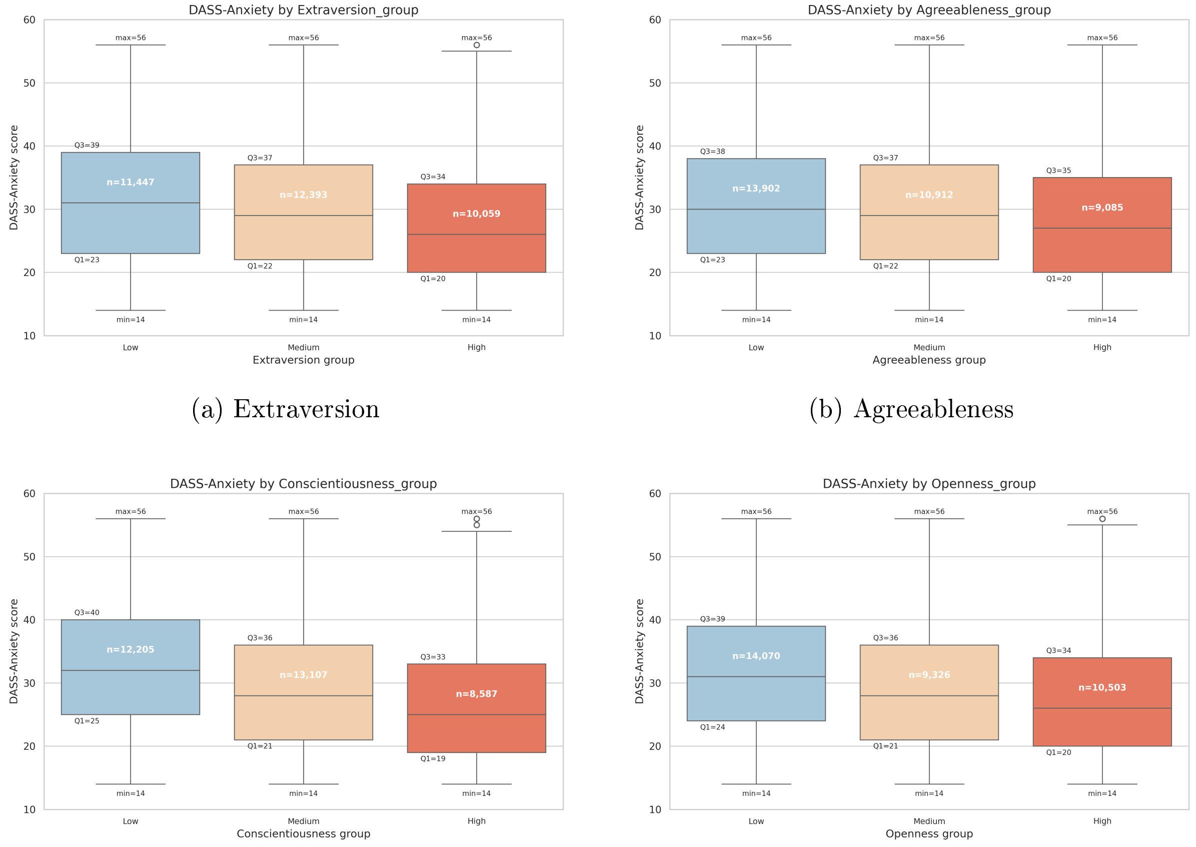


Figure 9: DASS anxiety scores by urban vs. rural living.

## A.2 Personality (TIPI) Boxplots

Figure ?? provides the full set of TIPI-based boxplots. While the main text focuses on Emotional Stability, the remaining traits (extraversion, agreeableness, conscientiousness, openness) display consistent but smaller gradients in anxiety scores.



(a) Extraversion

(b) Agreeableness

(c) Conscientiousness

(d) Openness

Figure 10: DASS anxiety scores across TIPI personality trait groups (low / medium / high). Emotional stability is shown in the main text.

## A.3 Nonresponse Diagnostics

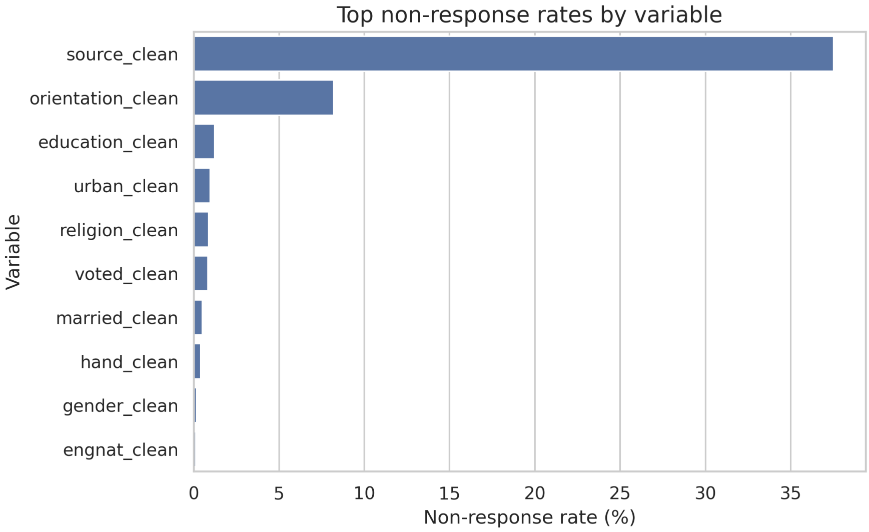


Figure 11: Nonresponse rates for the 10 most frequently skipped survey items.

## A.4 Additional Model Visualisations

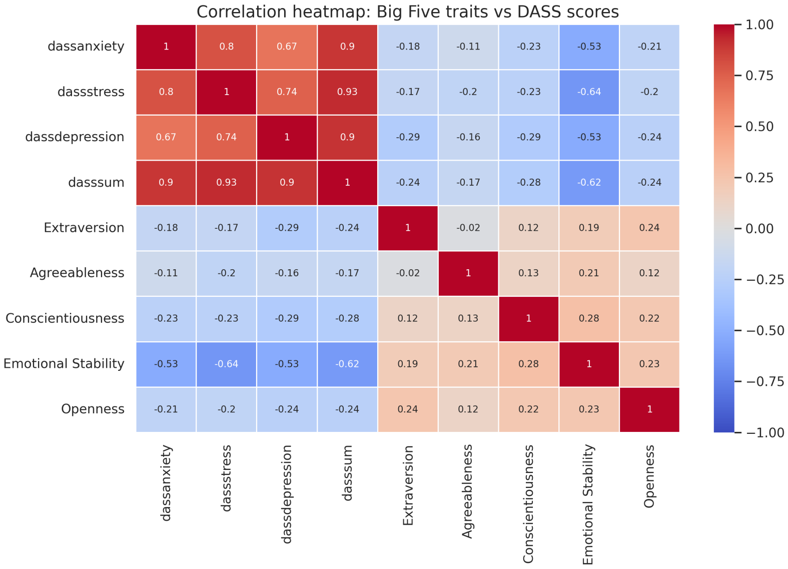


Figure 12: Heatmap of correlations between DASS scores and TIPI traits.

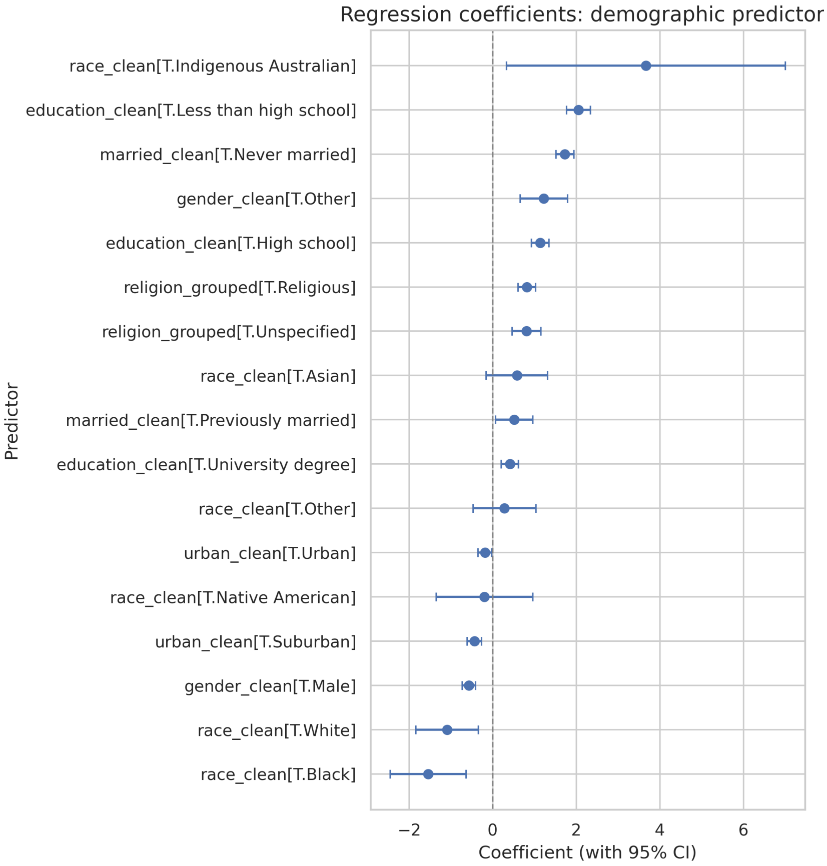


Figure 13: Linear regression coefficients for demographic predictors only.

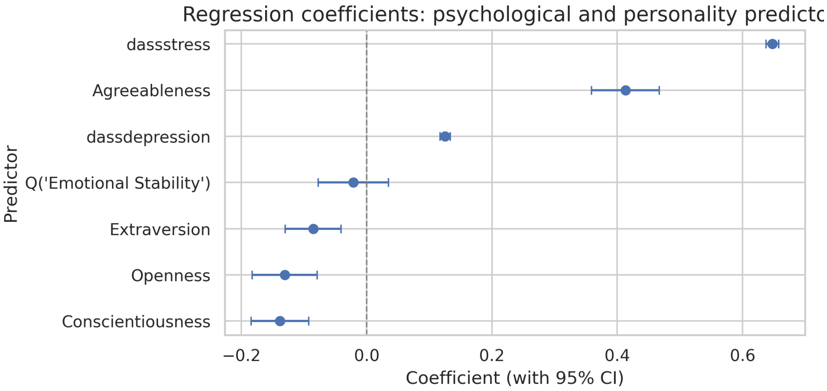


Figure 14: Linear regression coefficients for psychological and personality predictors.

### A.5 Additional Tables

Table **??** lists the ten countries with the highest average anxiety scores, complementing the continental overview given in the main results. The full ANOVA summaries for demographic and personality predictors are reported in Tables **??** and **??**.

Table 4: Top-10 countries by mean DASS anxiety score.

| Country | Mean anxiety |
|---|---:|
| Guam | 49.000000 |
| Kenya | 44.500000 |
| Ethiopia | 42.500000 |
| Senegal | 42.000000 |
| Isle of Man | 41.000000 |
| Suriname | 38.000000 |
| Kuwait | 37.285714 |
| Faroe Islands | 36.000000 |
| Maldives | 36.000000 |
| Montenegro | 35.333333 |

Table 5: Full ANOVA results for demographic predictors of anxiety.

| Factor | F | p-value |
|---|---:|---:|
| Gender | 342.909464 | 0.000000 |
| Sexual orientation | 162.446069 | 0.000000 |
| Marital status | 478.249891 | 0.000000 |
| Urban vs. rural | 36.984883 | 0.000000 |
| Race / ethnicity | 21.511400 | 0.000000 |
| Education | 330.595904 | 0.000000 |
| Religion | 29.350507 | 0.000000 |
| Age group | 538.873335 | 0.000000 |

Table 6: Full ANOVA results for personality (TIPI) predictors of anxiety.

| Trait | F | p-value |
|---|---:|---:|
| Extraversion | 434.725256 | 0.000000 |
| Agreeableness | 184.258496 | 0.000000 |
| Conscientiousness | 812.581633 | 0.000000 |
| Emotional Stability | 5205.264082 | 0.000000 |
| Openness | 577.828122 | 0.000000 |

Table 7: Comparison of regression models for predicting DASS-Anxiety.

| **Model** | $R^2$ **train** | $R^2$ **test** | **RMSE train** | **RMSE test** |
|---|---|---|---|---|
| Linear Regression | 0.672860 | 0.672006 | 5.776832 | 5.7669700 |
| Random Forest | 0.955513 | 0.670579 | 2.130289 | 5.7795067 |
| Gradient Boosting | 0.692757 | 0.683162 | 5.598395 | 5.668052 |

# B Code Repository

The full source code for this project is available in a public GitHub repository:

**GitHub Repository:** `https://github.com/vrelvape/mental-health-anxiety-dass21` The repository contains:

- A modular `src/` directory with cleaning, analysis, visualisation and modelling scripts.
- A `results/` directory with all generated figures and tables.
- A `tests/` directory with a small but focused test suite.
- Documentation files such as `README.md` and `AI_USAGE.md`.

All figures and tables shown in this report can be reproduced by running:

```
python -m src.main
```

from the project root, after installing the required dependencies.