Data Science and Advanced Programming 2025

Mental Health Anxiety

Final Project Report

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

December 8, 2025

Abstract

This project analyses how psychometric and demographic factors relate to self-reported anxiety, using data from the open DASS-21 survey (around 39,000 respondents). The aim is to describe how anxiety scores are distributed in the sample and to test which variables are most strongly associated with higher anxiety.

After cleaning and preprocessing the data, we summarise anxiety levels across age groups, genders, continents and countries, and compute basic descriptive statistics. We also produce a world map of mean anxiety by country and several boxplots to visualise differences between subgroups.

In a second step, we fit a set of regression models (linear regression, Random Forest and gradient boosting) to predict anxiety scores. Feature importances from the Random Forest are used to highlight the most influential predictors.

Overall, the analysis suggests that personality traits, especially low emotional stability, and co-occurring symptoms (stress and depression) are more strongly linked to anxiety than demographic variables. The entire workflow is implemented in Python and exports all figures and tables in a reproducible way.

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

1 Introduction

Anxiety disorders are among the most prevalent mental health conditions worldwide and can strongly affect daily functioning, work, and relationships. Yet, many individuals remain undiagnosed or untreated because screening is unevenly implemented and access to care is limited. Large-scale questionnaire data provide an opportunity to examine how anxiety relates to other psychological and demographic factors in a more systematic way.

In this report, I analyse an open DASS-21 (Depression Anxiety Stress Scales) dataset comprising more than 39,000 anonymous online responses. The data include psychometric scores for anxiety, depression, and stress, as well as basic demographic information such as age, gender, education, employment status, and country of residence.

The work has two main objectives. First, I describe and visualise how self-reported anxiety levels vary across demographic groups and world regions. Second, I build a predictive model that estimates DASS-21 anxiety scores from other variables, in order to identify which factors contribute most strongly to higher anxiety levels. Along the way, I also address practical issues such as missing data, imbalanced categories, and the interpretability of machine learning models.

More concretely, the analysis is guided by 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 organised as follows. Section 2 reviews related work and the theoretical context. Section 3 describes the dataset, preprocessing steps, and modelling approach. Section 4 presents the main empirical findings. Section 5 discusses the results, limitations, and ethical aspects. Finally, Section 6 summarises the work and outlines possible extensions.

2 Background and Related Work

The DASS-21 scale, introduced by Lovibond and Lovibond (1995), is a widely used instrument to measure depression, anxiety and stress. Large DASS-based datasets have been used to study psychological well-being in different populations, showing that anxiety often varies with age, gender, employment status and other sociodemographic factors, and that it tends to co-occur with stress and depressive symptoms.

Recent work in data science has applied machine learning models such as linear and logistic regression, decision trees and random forests to predict mental health outcomes from survey data. These models can achieve good predictive performance but raise questions about interpretability and fairness, especially when predictions are linked to sensitive characteristics. Some studies therefore complement predictive models with feature-importance measures or post-hoc explanation tools.

This project follows that line of work but focuses on a single openly available DASS-21 dataset. The goal is to combine descriptive statistics and relatively simple predictive models in order to (i) document how anxiety is distributed across groups in this sample and (ii) quantify the relative importance of psychological, personality and demographic predictors, while keeping in mind the limitations of an online convenience sample.

3 Methodology and Implementation

3.1 Data Description

We use the open-access **DASS-21 survey** dataset from Open Psychometrics, which contains a little over 39,000 anonymised online responses. The questionnaire measures Depression, Anxiety

and Stress via 21 Likert-scale items, and is complemented by a set of self-reported demographic and personality variables.
In this project we mainly rely on:

- **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).

Although the dataset is large, it has the usual problems of online surveys: missing values, heterogeneous free-text country labels, and unbalanced group sizes (for example, many more female respondents than male participants). The sample is also **geographically skewed**: a substantial fraction of responses comes from a small number of countries (notably Malaysia), whereas many other countries have only a few participants. As a result, cross-country and continental comparisons later in the report should be interpreted with caution and not as representative estimates for the general population.

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 standardised 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 broader 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.

The full preprocessing pipeline can be re-run from scratch, which makes the analyses in the following sections fully reproducible.

- **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 with `pycountry`; religion and orientation grouped into a smaller set of categories; creation of age bands).
- **Handling missing and nonresponse**: rows with missing values in key variables are dropped for most analyses, and explicit `"No response"` categories are removed from the predictive models to avoid mixing observed values with nonresponse.

3.3 Nonresponse Bias and Data Quality

As already noted in the project proposal feedback, survey data can be affected by **nonresponse bias**. In this dataset, several items related to gender identity and sexual orientation have noticeably higher nonresponse rates (above 10%). To keep the main text readable, a detailed analysis of the most frequently skipped items is reported in the Appendix (Figure 11).

Beyond item-level nonresponse, the sample is also **imbalanced by gender**, with substantially more female than male respondents, and **geographically skewed**, with a concentration of participants in a small number of countries. If left as is, these imbalances can influence both descriptive statistics and predictive models. Here, two pragmatic choices were made:

- Nonresponse values and explicit `"No response"` categories are excluded from the Random Forest model, so that predictions are based only on observed information.
- The Random Forest is trained with **inverse-frequency sample weights** for gender, giving more weight to under-represented groups (e.g., male respondents) and reducing the impact of the raw imbalance on the model.

We do *not* attempt to reweight by country, so geographic skew remains a limitation for external validity. These decisions do not remove bias completely, but they reduce the dependence of the model on the most over-represented subgroups. Remaining limitations are discussed in Section 5.

3.4 Modelling and Implementation

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

- `src/cleaning.py`: cleaning and harmonisation pipeline that produces the analysis dataframes.
- `src/analysis/`: statistical analyses such as ANOVAs by demographics and personality.
- `src/viz/`: visualisations (histograms, boxplots, maps, forest plots).
- `src/analysis/ml_models.py`: regression models for anxiety (linear, Random Forest, gradient boosting) and model comparison utilities.
- `src/main.py`: script that runs the end-to-end pipeline and exports 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 differences are examined using one-way ANOVAs for both demographic factors (e.g., gender, age group, marital status) and personality traits (grouped into low, medium and high tertiles).

For the predictive part, anxiety is treated as a **continuous outcome**. Three supervised algorithms are compared 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 removed before fitting. Performance is evaluated on a hold-out test set using the coefficient of determination (R^2) and root mean squared error (RMSE). For model interpretation, we focus on the Random Forest and report global feature importance scores, exported as both a table and a barplot.

3.5 Testing and Reproducibility

To make the analysis reproducible, the project is organised as a simple but fully scripted pipeline:

- Running `python -m src.main` launches the full workflow: dataloading, cleaning, statistical analyses, visualisations and machine learning, and saves outputs under `results/figures` and `results/tables`.
- A small 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.

With this structure, all figures and tables in the report can be regenerated from the raw dataset with a single command, while basic automated checks help detect obvious issues in the main components of the codebse.

4 Results and Evaluation

4.1 Overall Anxiety Levels

To get a first overview of anxiety in the sample, we examined the distribution of DASS-Anxiety scores using a boxplot with clinical severity bands (Figure 1). 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 respondents in the Normal category. The distribution is therefore shifted towards higher severity levels, suggesting that this online sample reports substantially elevated anxiety compared to what would be expected in a general population.

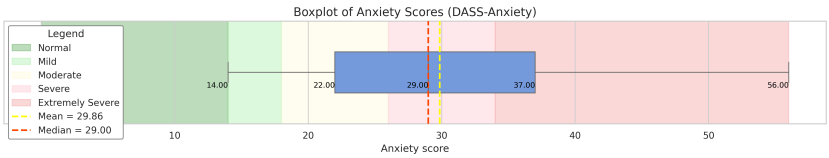


Figure 1: Distribution of DASS-Anxiety scores with clinical severity thresholds.

The corresponding summary table (Table 1) 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 are particularly pronounced. Figure 2 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).

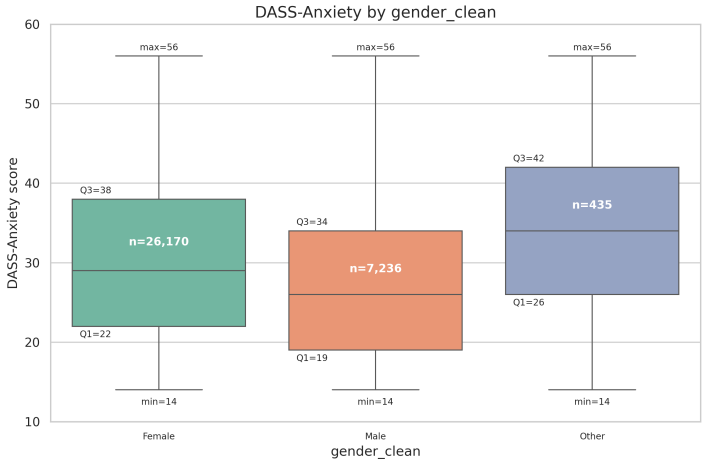


Figure 2: Boxplot of DASS-Anxiety scores by gender. Median anxiety is higher in female respondents than in male respondents.

Geographical analyses revealed smaller but systematic differences. Table 2 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), which suggests 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 how large-scale survey data can highlight populations where anxiety may be particularly elevated.

4.3 Personality and Anxiety

We then examined how Big Five personality traits, as measured by the TIPI-10 scale, relate to anxiety. Figure 3 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.

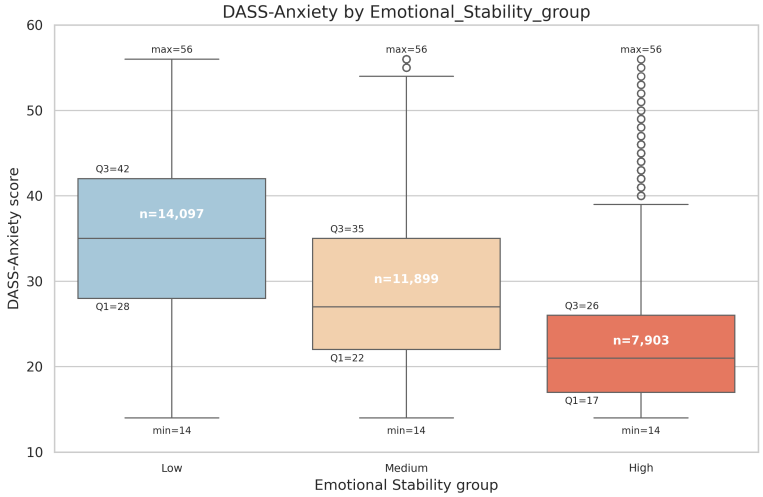


Figure 3: DASS-Anxiety scores by Emotional Stability group (TIPI). Lower Emotional Stability is associated with markedly higher anxiety.

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 3 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 fit well with established psychological theory, where high neuroticism (low emotional stability) is known to be a strong risk factor for anxiety. In the next 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

We compared three regression models (linear regression, Random Forest and Gradient Boosting) 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, sexual 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 to make the analysis reproducible.

One practical issue is the unbalanced gender distribution: there are many more female than male respondents in the sample. To reduce the impact of this imbalance, the Random Forest was trained with *inverse-frequency sample weights* by gender, so that under-represented groups

received greater weight during training. This does not remove bias entirely, but it limits the extent to which the model is driven by the majority group.

Appendix Table 7 summarises the performance of the three models. The linear regression baseline has 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. This gap indicates some overfitting, but the test performance still shows that the model captures a large share of the variance in anxiety scores.

Figure 4 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 the other TIPI traits. Demographic variables such as age group, gender and education also contribute, but with smaller importance values. Overall, this suggests that co-occurring symptoms and personality factors are more informative for predicting anxiety than sociodemographic characteristics alone.

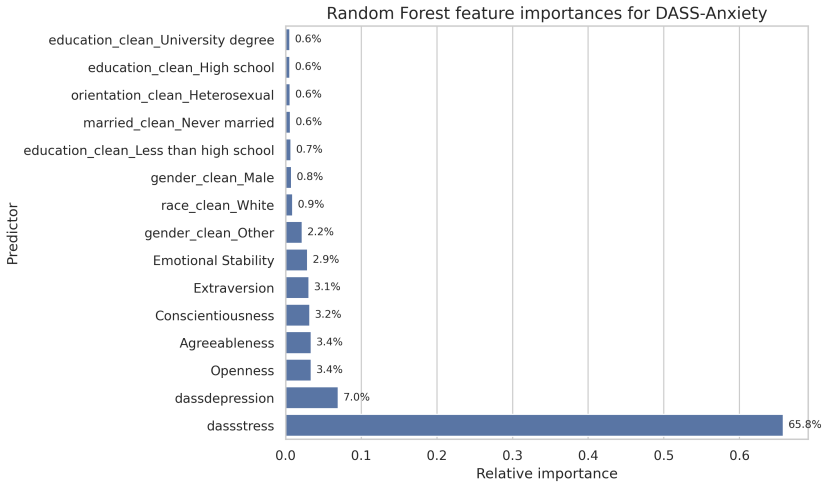


Figure 4: Top 15 predictors of DASS-Anxiety according to the Random Forest regression model (with inverse-frequency gender weighting).

Taken together, 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.

5 Discussion

The goal of this project was to characterise anxiety levels in a large online DASS-21 sample and to explore how demographic and psychological variables relate to self-reported anxiety. In our data, anxiety scores are generally high: more than half of respondents fall in the moderate or higher range. Differences between groups (e.g., by gender, age or personality profile) are present and systematic, but they are not dramatic in absolute terms.

What worked well. One practical strength of the project is the fully reproducible pipeline, which links 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 pipeline produced analysis-ready tables and a coherent set of figures.

The main descriptive results also fit together in a sensible way. The severity distribution (Figure 1) shows that anxiety is shifted towards higher categories, and this pattern reappears in subgroup analyses by gender, age, continent and other demographic factors. The personality analyses and the Random Forest model both highlight emotional stability, stress and depression as key variables, which increases confidence that these are not just random findings.

Comparison with expectations. Several patterns match what one would expect from the literature. Women report higher anxiety than men (Figure 2); younger adults tend to be more anxious than older groups; and urban respondents show slightly higher scores than rural participants. These trends are in line with typical epidemiological results and probably reflect a mix of social stressors, help-seeking behaviour and gendered norms around emotional expression. On the psychological side, the strong negative association between emotional stability and anxiety (Figure 3; Table 3) is consistent with Big Five theory. The Random Forest results, which give very high importance to `dassstress` and `dassdepression`, also fit with the known comorbidity between anxiety, stress and depressive symptoms.

Challenges and sources of bias. At the same time, several aspects of the dataset limit how far the results can be generalised. First, the data come from an online convenience sample rather than a representative survey. 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 is clearly not random: items about gender identity and sexual orientation have much higher nonresponse rates than most other variables, which can bias subgroup comparisons. We tried to mitigate these issues by (i) removing "No response" categories when preparing the modelling dataset and (ii) 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 led to feature rankings that were very similar to the unweighted model, which suggests that the main psychological signals (stress, depression, emotional stability) are not purely artefacts of gender imbalance. However, some residual bias almost certainly remains, especially for smaller groups such as non-heterosexual or non-binary participants.

Interpretation of effect sizes. Because the sample is very large, many group differences are statistically significant (ANOVA $p < .001$ in most cases), but the associated effect sizes are often modest. For example, mean anxiety scores differ by only one to two points between continents (Table 2), 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 relatively good for psychological data but still far from deterministic. Overall, the results point towards anxiety being influenced by a broad combination of factors, many of which are not observed in this dataset (e.g., life events, socioeconomic status, physical health, access to care). The model is therefore useful for describing 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 are worth noting. The sample is geographically concentrated in a handful of countries (especially Malaysia), so the world map and country-level comparisons should be seen as patterns within this particular dataset rather than as global estimates. All measures are self-reported and collected at a single time point. This makes the results vulnerable to response styles and momentary mood, and it rules out strong causal interpretations. The DASS subscales are also partly overlapping by design, so the very high importance of `dassstress` and `dassdepression` in the Random Forest is not surprising: 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 conceal more nuanced patterns. Finally, the modelling work focused on a small set of algorithms (linear regression, Random Forest and Gradient Boosting) and did not explore more advanced approaches such as calibrated probabilistic models or hierarchical structures. This was a deliberate choice for a first project, but it leaves room for more sophisticated modelling in future work.

Additional analyses in the Appendix. The Appendix includes several additional analyses that broadly support the main findings. Demographic boxplots (by age group, education, race, religion, marital status, urbanicity and sexual orientation) show similar trends: 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 reinforce 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 many group differences are statistically significant, largely because of the sample size, but that effect sizes remain modest. Taken together, these supplementary results are consistent with the main message of the study: anxiety is widely elevated in this sample, with systematic but not extreme differences across demographic and psychological profiles.

6 Conclusion and Future Work

6.1 Summary

In this project, I analysed more than 39,000 responses to the DASS-21 questionnaire using a full data science pipeline to better understand and predict anxiety scores. The workflow cleans the raw data, harmonises categorical variables, builds 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 and comparatively few in the normal band. 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.

The tree-based ensemble models (Random Forest and gradient boosting) achieved 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 were psychological: stress and depression scores, together with low emotional stability, consistently emerged as the strongest correlates. Personality traits such as conscientiousness and extraversion also contributed, whereas demographic variables added relatively little predictive power once these psychological factors were taken into account.

Taken together, these findings suggest that, in this dataset, anxiety is more strongly related to co-occurring symptoms and stable personality dimensions than to basic sociodemographic characteristics.

6.2 Future Directions

Several extensions would be worth exploring. On the data side, studies using more balanced and deliberately sampled populations (for example stratified by gender, age and region) or longitudinal designs would make it easier to generalise and to track changes in anxiety over time. Methodologically, future work could compare alternative strategies for handling class imbalance, test additional algorithms, and incorporate richer socioeconomic or clinical variables. It would also be useful to evaluate model performance and fairness explicitly across key demographic subgroups (e.g. gender, age, region), rather than only reporting aggregate metrics.

References

[1] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[2] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. "A Very Brief Measure of the Big-Five Personality Domains". In: *Journal of Research in Personality* 37.6 (2003), pp. 504–528.

[3] Peter F. Lovibond and Sydney H. Lovibond. *Manual for the Depression Anxiety Stress Scales*. 2nd ed. Sydney: Psychology Foundation of Australia, 1995.

[4] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. 2017.

[5] Open Psychometrics. *Depression Anxiety Stress Scales (DASS-21) Raw Data*. Online dataset. URL: `https://openpsychometrics.org/_rawdata/` (visited on 11/10/2025).

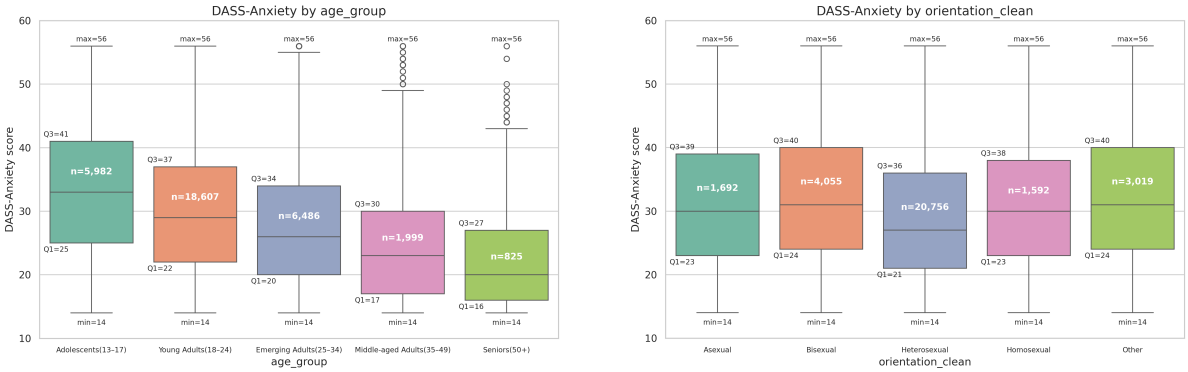
[6] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

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 4 and Section 5.

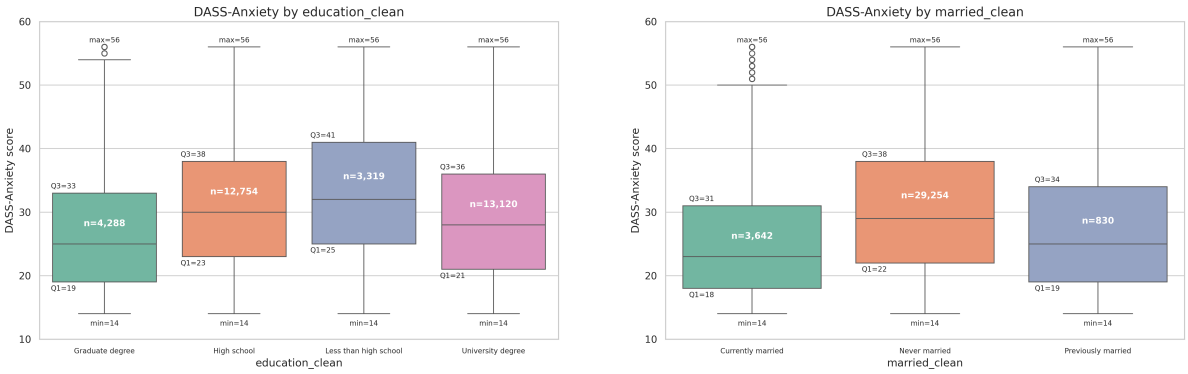
A.1 Demographic Boxplots

Figure 5 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.

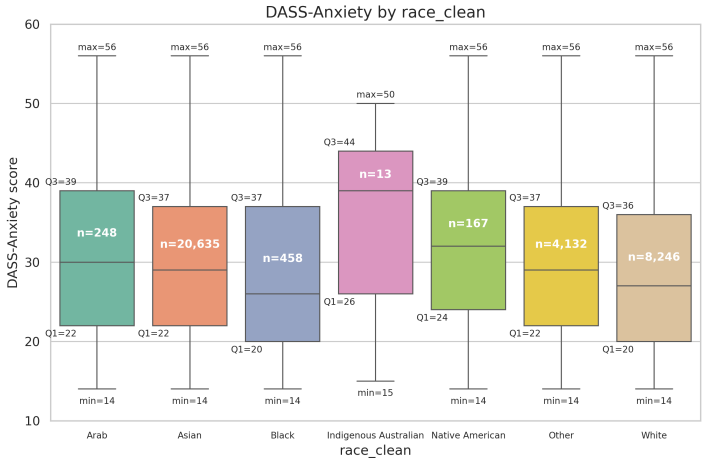


Figure 7: DASS anxiety scores by race/ethnicity.

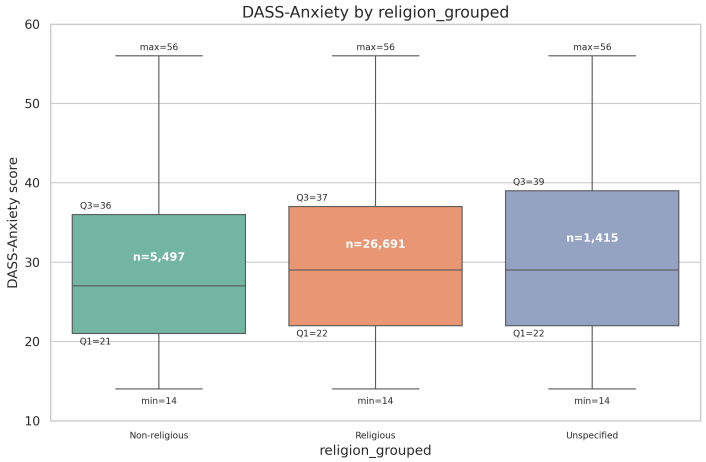


Figure 8: DASS anxiety scores by religion (grouped categories).

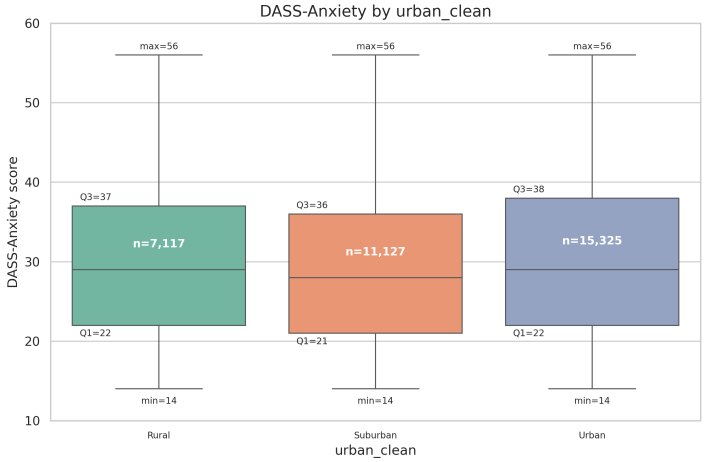
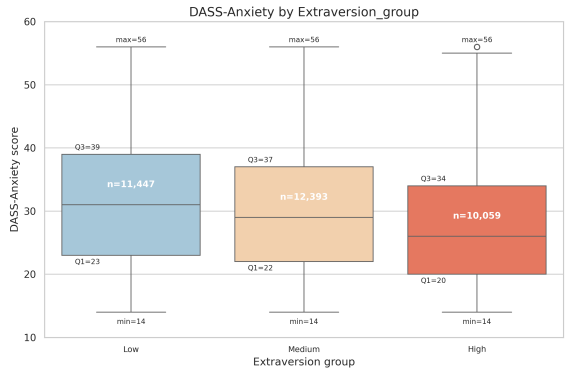


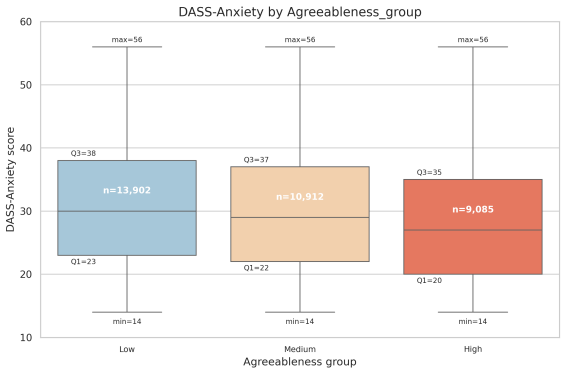
Figure 9: DASS anxiety scores by urban vs. rural living.

A.2 Personality (TIPI) Boxplots

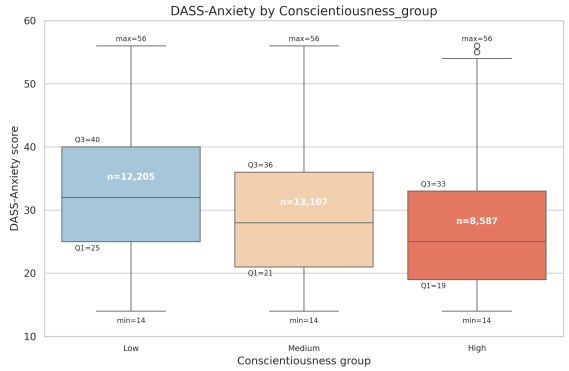
Figure 10 provides the full set of TIPI-based boxplots. While the main text focuses on Emotional Stability, the remaining traits (extraversion, agreeableness, conscientiousness, openness) display consistent but smaller gradients in anxiety scores.



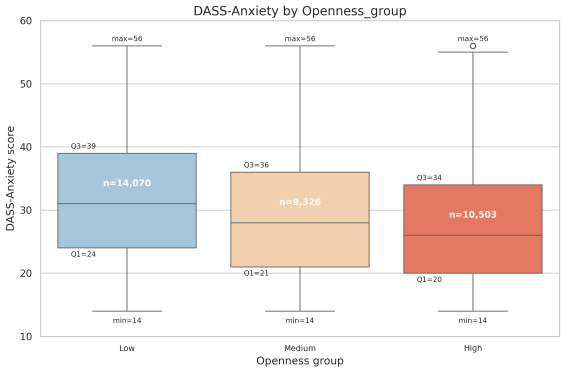
(a) Extraversion



(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

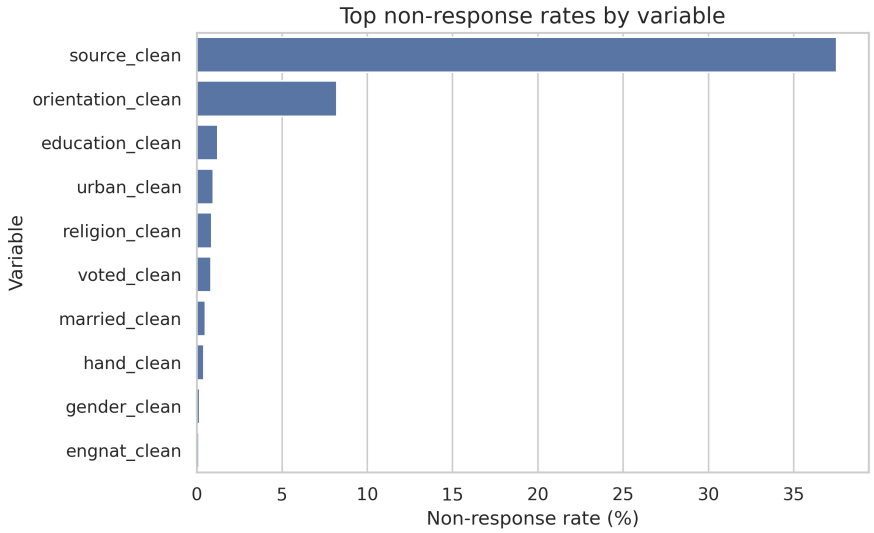


Figure 11: Nonresponse rates for the 10 most frequently skipped survey items.

A.4 Additional Model Visualisations

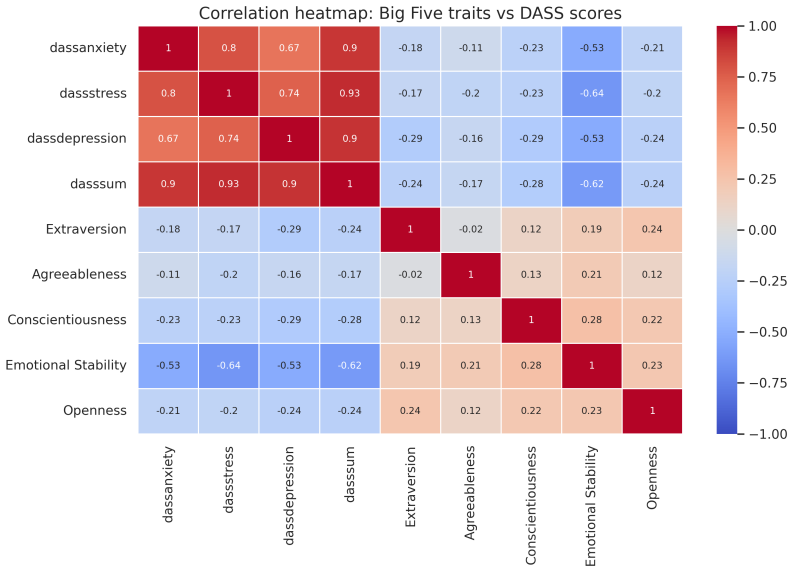


Figure 12: Heatmap of correlations between DASS scores and TIPI traits.

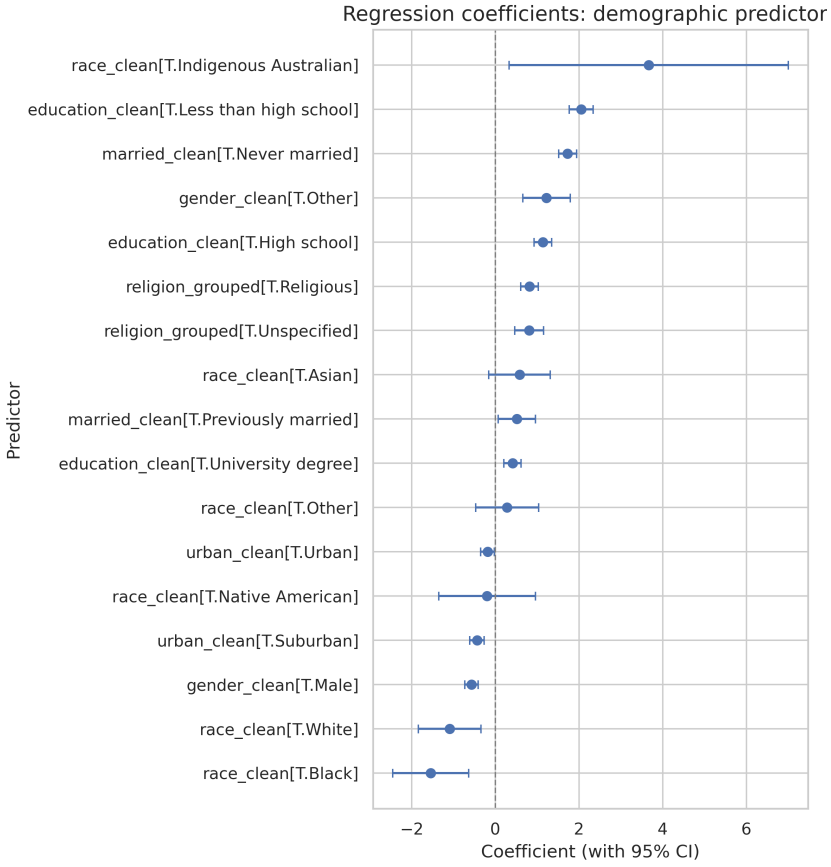


Figure 13: Linear regression coefficients for demographic predictors only.

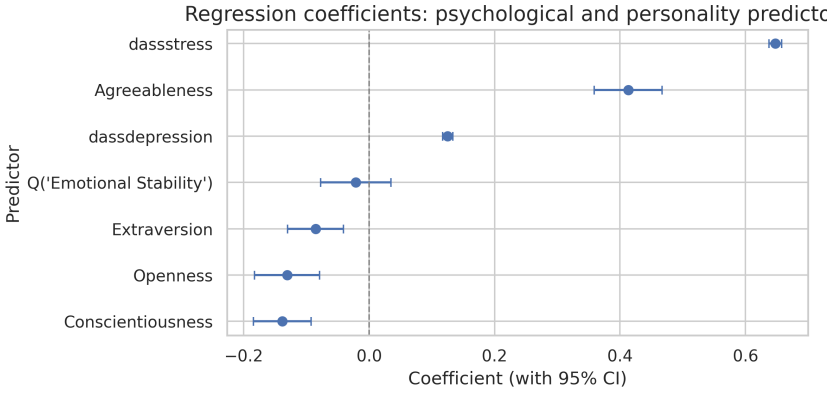


Figure 14: Linear regression coefficients for psychological and personality predictors.

A.5 Additional Tables

Table 4 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 5 and 6.

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.

C Use of AI-based Helper Tools

In line with the Advanced Programming 2025 project guidelines, this appendix briefly reports how AI-based helper tools were used during the development of this project.

AI tools used

- **ChatGPT** (OpenAI) – used as a conversational assistant for programming support, LaTeX help, and feedback on the structure and wording of the report.

No other AI coding assistants (such as GitHub Copilot) were used.

Scope of assistance

AI support was used in a limited and targeted way:

- **Debugging and code refinement**: Getting suggestions to understand and fix Python errors and warnings, simplify functions, and make the data-processing pipeline more robust (e.g., handling missing values, cleaning categorical variables, and structuring the code into reusable components).
- **Library usage and implementation details**: Clarifying how to use specific features of `pandas`, `scikit-learn`, and plotting libraries (for example, configuring a `RandomForestRegressor`, computing and visualising feature importances, or producing clear boxplots and heatmaps).
- **LaTeX and typesetting**: Help with resolving LaTeX compilation issues (missing packages, misformatted tables, undefined references) and with converting Python outputs (tables and figures) into LaTeX environments using `booktabs`, captions, and labels.
- **Writing and structure**: High-level suggestions for how to organise the Introduction, Methodology, Results, Discussion, and Conclusion sections, and for improving the clarity of some passages. All text was read, edited, and adapted by the author to reflect their own understanding and style.
- **Conceptual checks**: Explanations and sanity checks for statistical and machine learning concepts (e.g., ANOVA, effect sizes, overfitting, train/test splits, interpretation of R^2 and feature importances) to make sure that the chosen methods and interpretations are reasonable.

Limitations of AI assistance and author responsibility

AI tools were used strictly as helpers, not as a replacement for the author's work:

- All code in the repository was written, adapted, or refactored by the author and executed locally to verify that it runs and reproduces the results.
- All data-processing choices, modelling decisions, and interpretations of the results are the responsibility of the author. AI suggestions were only adopted after being checked and, when necessary, modified.
- The final structure of the report, the selection of figures and tables, and the main conclusions are based on the empirical results and on the author's judgement.

This appendix is intended to provide transparent documentation of how AI-based tools were used, in accordance with the course policy.