

Homework 2: Principal Component Analysis

S&DS 5360 | Multivariate Statistics

Brandon Tran (bat53)

January 30, 2026

```
library(tidyverse)
library(haven)
```

Setup & Data Overview

Over the course of the semester, we will be using the Healthy Minds Network’s latest 2024–2025 Healthy Minds Study (HMS) student survey. The HMS is a population-level, web-based survey administered by the HMN to assess mental health, service utilization, and related factors among post-secondary students. Since its inception over 15 years ago, the study has collected more than 935,000 responses from over 675 institutions nationwide.

The latest 2024–2025 iteration was graciously made available to us following a successful application for access, which proposed the following research questions (framed in the context of S&DS 5360’s curriculum):

The proposed research aims to examine the multivariate relationship between sexual and gender minority (GSM) identities, digital environment exposure, and mental health outcomes in the 2024-25 college student population. Utilizing the Demographics and Mental Health Status standard modules, this study will investigate how latent factors of psychological distress—comprising the PHQ-9 (Depression), GAD-7 (Anxiety), and UCLA Loneliness Scale—vary across intersectional identities. Additional exploratory analysis may include (1) factor analysis / PCA to identify if “Psychological Distress” and “Academic Impairment” (e.g., deprawsc, anx_score, and aca_impact) load onto distinct latent constructs for different GSM generations, (2) MANOVA to determine if the “vector of well-being” (comprising flourishing, loneliness, and distress scores) differs significantly based on the interaction between categorical identity markers and experiences of unfair treatment at the institutional or digital level, and (3) cluster analysis to identify “Student

Behavioral Archetypes” based on a combination of lifestyle factors (sleep_wknight, exerc) and digital engagement.

The resulting dataset features $n = 84,735$ individual respondents across 1,608 questions. These questions are organized into three module types:

- **Standard Modules:** Fielded at all institutions and provide core data on demographics, mental health status (e.g., depression and anxiety screenings), and help-seeking behaviors.
- **Elective Modules:** Specific topics requested by select institutions for in-depth assessment (e.g., substance use, eating and body image, AI attitudes).
- **Special Modules:** Other unique modules tailored for specific cohorts (e.g., Black College Student Mental Health Module for HBCUs).

To ensure that the results are representative of the broader student population at each school, the dataset includes a non-response weight variable, `nrweight`. It adjusts for the fact that female students typically respond at higher rates than male students, and it gives equal aggregate weight to each school in the national estimates, preventing results from being dominated by very large institutions.

For the purpose of Homework 2, which focuses on Principal Component Analysis, we will focus on questions that fall under the Standard Modules. Specifically, we will use composite variables that are created during cleaning and are the sum of a series of Likert scale questions on the same topic. The composite variables are defined as the sum of each nested question:

- **flourish:** Below are 8 statements with which you may agree or disagree. Using the 1-7 scale, indicate your agreement with each item by indicating that response for each statement. ($min = 8$, $max = 56$)
 1. I lead a purposeful and meaningful life.
 2. My social relationships are supportive and rewarding.
 3. My social relationships are supportive and rewarding.
 4. I am engaged and interested in my daily activities.
 5. I actively contribute to the happiness and well-being of others.
 6. I am competent and capable in the activities that are important to me.
 7. I am a good person and live a good life.
 8. I am optimistic about my future.
- **dpraws:** Over the last 2 weeks, how often have you been bothered by any of the following problems? ($min = 0$, $max = 27$)
 1. Little interest or pleasure in doing things
 2. Feeling down, depressed or hopeless
 3. Trouble falling or staying asleep, or sleeping too much
 4. Feeling tired or having little energy
 5. Poor appetite or overeating

6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down
 7. Trouble concentrating on things, such as reading the newspaper or watching television
 8. Moving or speaking so slowly that other people could have noticed; or the opposite—being so fidgety or restless that you have been moving around a lot more than usual
 9. Thoughts that you would be better off dead or of hurting yourself in some way
- **anx_score:** Over the last 2 weeks, how often have you been bothered by the following problems? (*min* = 0, *max* = 21)
 1. Feeling nervous, anxious or on edge
 2. Not being able to stop or control worrying
 3. Worrying too much about different things
 4. Trouble relaxing Becoming easily annoyed or irritable
 5. Being so restless that it's hard to sit still
 6. Feeling afraid as if something awful might happen
 - **lonesc:** Please answer the following: (*min* = 3, *max* = 9)
 1. How often do you feel that you lack companionship?
 2. How often do you feel left out?
 3. How often do you feel isolated from others?
 - **ed_sde:** Please answer the following questions as honestly as possible. (*min* = 0, *max* = 5)
 1. Do you often feel the desire to eat when you are emotionally upset or stressed?
 2. Do you often feel that you can't control what or how much you eat?
 3. Do you sometimes make yourself throw up (vomit) to control your weight?
 4. Are you often preoccupied with a desire to be thinner?
 5. Do you believe yourself to be fat when others say you are thin?

Omitting NAs, we fall from 84,735 to 68,062 unique responses to each of the above modules. With our variables defined, we proceed to load in our data for subsequent analyses.

```
hms_data <- read_sav("../data/hms_data.sav", encoding = 'latin1')
```

```
pca_data <- hms_data %>%
  select(
    flourish,
    deprawsc,
    anx_score,
    lonesc,
    ed_sde
```

```
) %>%  
drop_na()
```