

# Homework 2: Principal Component Analysis

S&DS 5360 | Multivariate Statistics

Brandon Tran (bat53)

January 31, 2026

```
library(tidyverse)
library(knitr)
library(haven)
library(corrplot)
library(PerformanceAnalytics)
library(heplots)
source("https://raw.githubusercontent.com/jreuning/sds363_code/refs/heads/main/pcaThreshold.r")
source("https://raw.githubusercontent.com/jreuning/sds363_code/refs/heads/main/ciscoreplot.r")
```

## 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 institution 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\_impa) 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.
- **dprawsc:** 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 relaxingBecoming 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()

head(pca_data) # Including for grader to easily orient to data :)
```

```

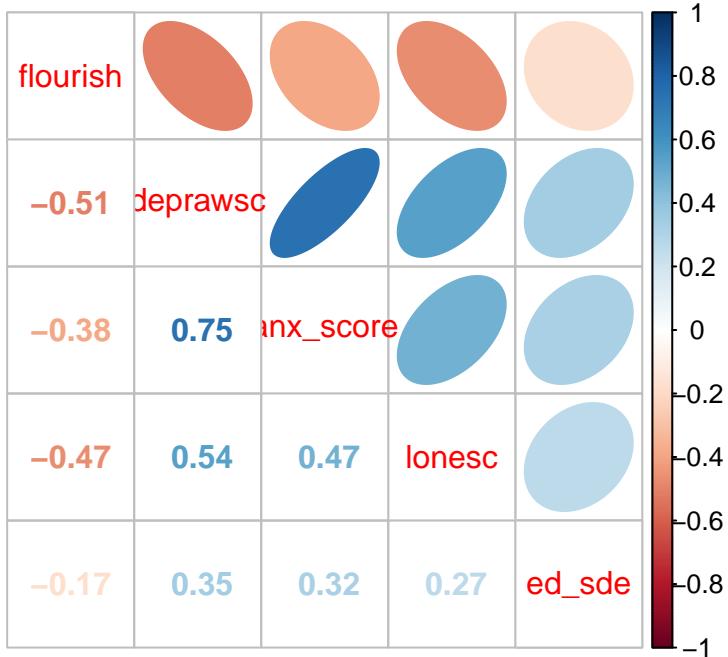
# A tibble: 6 x 5
  flourish deprawsc anx_score lonesc ed_sde
  <dbl>     <dbl>      <dbl>    <dbl>    <dbl>
1     26      24        13      7      2
2     32      8         7       6      1
3     47      8         9       6      3
4     48      6         6       5      0
5     28      7        14      9      2
6     41      7         8       7      3
```

## Part 1 | Exploratory Data Analysis

### High-Dimensional Linearity

```

corr_matrix <- cor(pca_data)
corrplot.mixed(corr_matrix, lower = "number", upper = "ellipse")
```

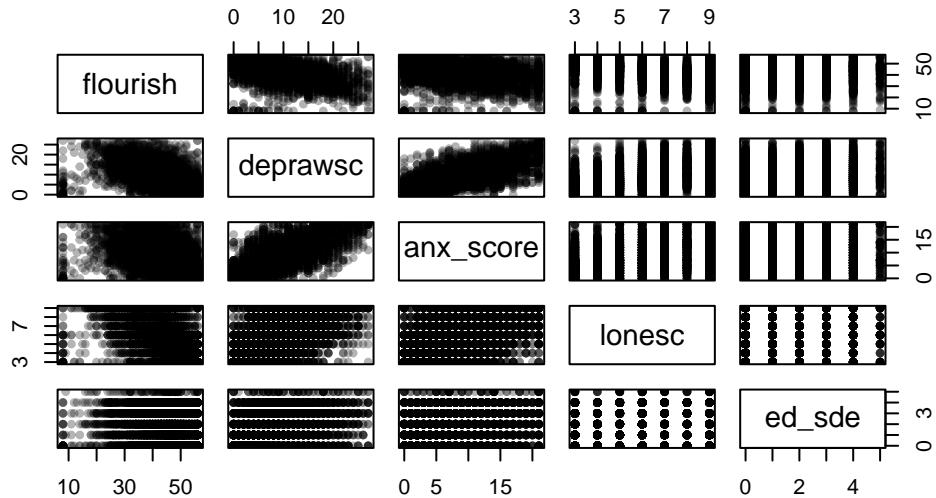


Our initial correlation plot suggests strong linear relationships between our composite variables. Observe the dark reds and blues: there appears to be a distinct cluster of positive correlations between depression (`deprawsc`), anxiety (`anx_score`), and loneliness (`lonesc`), which are all strongly negatively coordinated with flourishing (`flourish`). The strength of these coefficients (generally,  $|r| > 0.5$ ) suggests that the data exhibits linearity, making it a strong candidate for PCA. However, we will create a matrix plot to specifically confirm high-dimensional linearity.

```
pca_plot_sample <- pca_data %>%
  slice_sample(n = 5000)

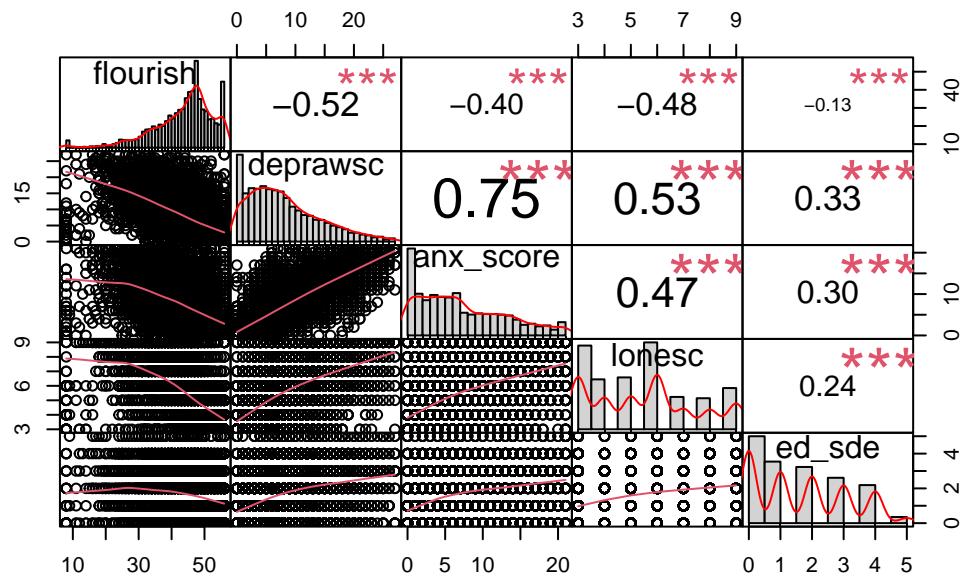
plot(
  pca_plot_sample,
  pch = 19,
  cex = 0.7,
  col = alpha("black", 0.3),
  main = "Matrix Plot of Sampled Composite Response Data")
```

## Matrix Plot of Sampled Composite Response Data



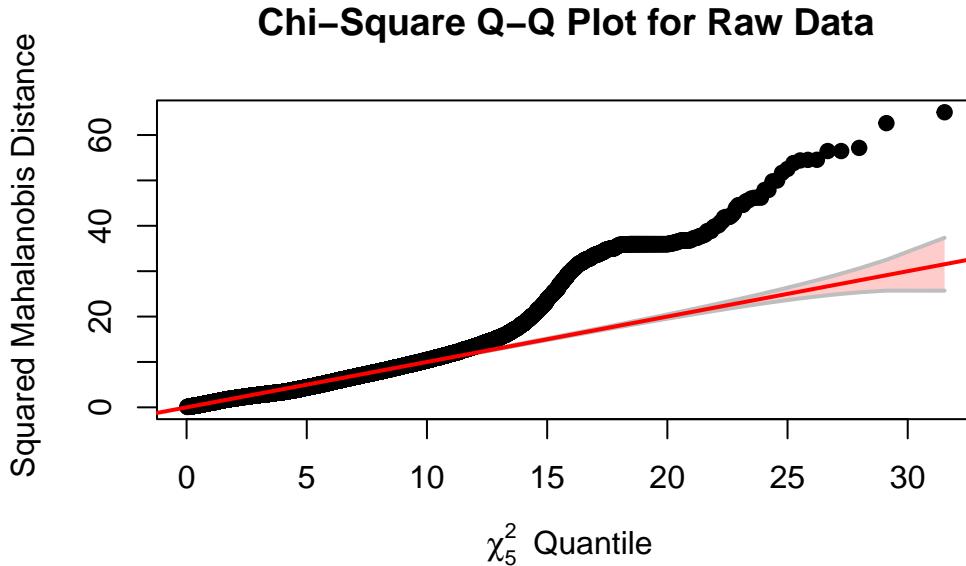
We sample 5,000 observations for our matrix plot, as plotting all 68,062 observations would result in overplotting and be computationally intensive. Immediately, we see the same strong linear correlations that we saw in the correlation plots: a positive correlation between depression and anxiety and a negative correlation between flourishing versus depression and anxiety. This is unsurprising. However, it is harder to see a clear relationship amongst `lonesc` and `ed_sde`. The grid-like striations do not necessarily reflect non-linearity (just discreteness), but we can further analyze with the correlation chart capabilities of the `PerformanceAnalytics` package.

```
chart.Correlation(pca_plot_sample)
```



The trends we see across flourishing, depression, anxiety, and loneliness remain consistent with earlier analyses, with added specificity on statistically significant relationships ( $p < 0.001$ ). The relationships between `flourish` and both `deprawsc` and `anx_score` appear to be slightly nonlinear when viewed via LOESS curves, whereas the relationship between `deprawsc` and `anx_score` remains highly linear. The LOESS curves for relationships with `ed_sde` and `lonesc` appear roughly linear, though with a potential quadratic component. Before considering transformations, we test for multivariate normality.

```
cqplot(pca_data, main = "Chi-Square Q-Q Plot for Raw Data")
```



We see strong structure and skewness in the Q-Q plot, suggesting that the data are not multivariate normal. At the cost of interpretability, we transform the variables, not only to attain multivariate normality (which is not required for PCA), but also to hopefully improve linearity and reduce outlier influence.

## Data Transformations

Given the output in our `PerformanceAnalytics` plot, we propose the following transformations:

- **Square Root:** `deprawsc` and `anx_score` to correct for right-skew
- **Power Transformation ( $x^2$ ):** `flourish` to correct for left-skew

Since `lonesc` and `ed_sde` have fewer levels, we do not propose transformations for these variables. Note that for the variables we take the square root of, we first increment by 0.5 to stabilize the variance given the plethora of zeroes in our data.

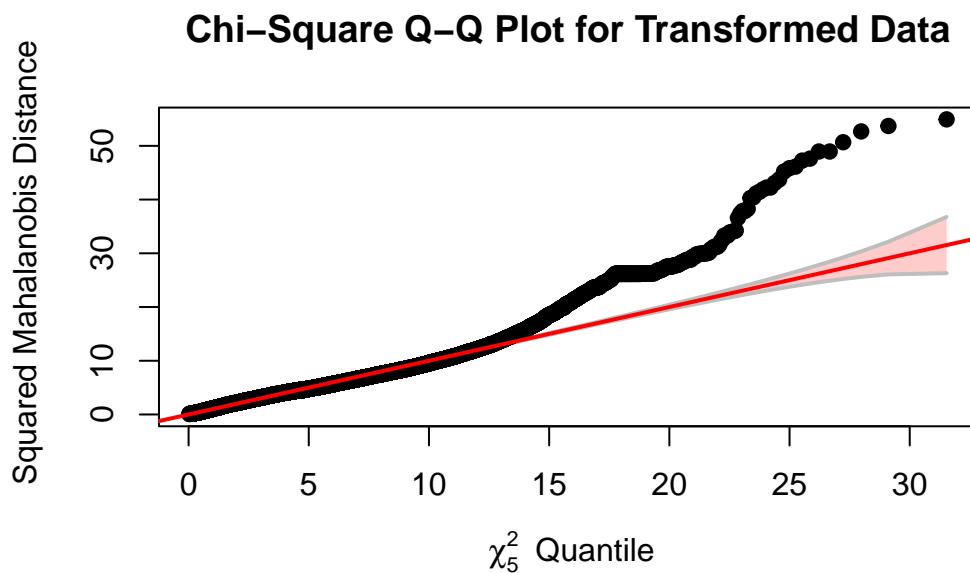
```
pca_data_transform <- pca_data %>%
  mutate(
    deprawsc_sqrt = sqrt(deprawsc + 0.5),
    anx_sqrt = sqrt(anx_score + 0.5),
    flourish_sq = flourish^2
  ) %>%
```

```
select(fLOURISH_sq, deprawsc_sqrt, anx_sqrt, lonesc, ed_sde)
```

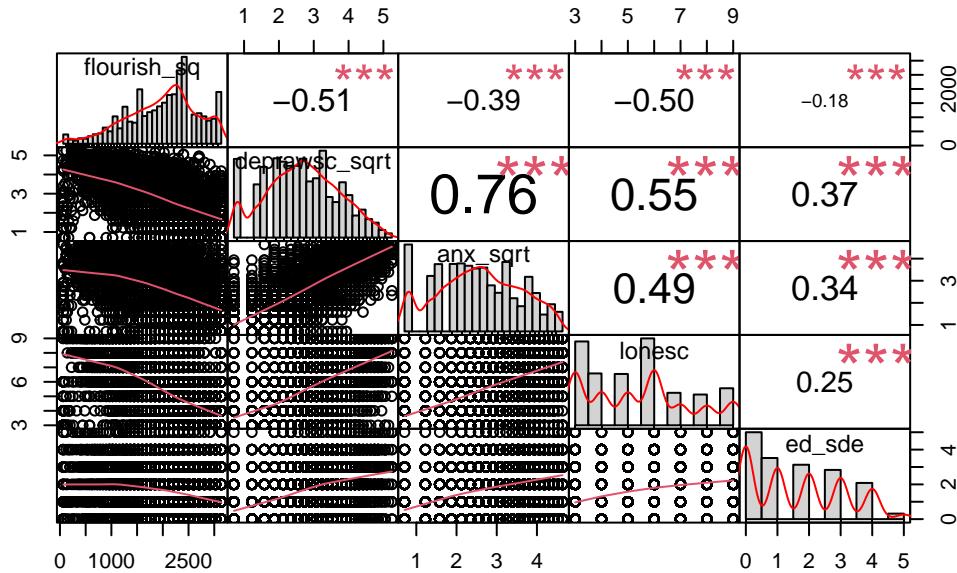
We review the same Chi-Square Q-Q and correlation plots for our transformed data.

## Multivariate Normality

```
cqplot(pca_data_transform, main = "Chi-Square Q-Q Plot for Transformed Data")
```



```
chart.Correlation(slice_sample(pca_data_transform, n = 5000))
```



While our data are still not multivariate normal, we notice much clearer linear trends in the scatterplots. We feel comfortable proceeding with PCA given our large sample size, high correlations, and high-dimensional linearity.

As an aside, we also considered a log transformation of `deprawsc` and `anx_score` but found that this introduced new artifacts due to the discrete nature of the data and generally over-corrected.

## Part 2 | Correlations

We presented a variety of correlation plots in Part 1 as part of our EDA but reproduce key visualizations here for post-transformation analysis. To begin, we present a side-by-side comparison of correlations by variable pair before and after transformations.

```
corr_matrix_2 <- cor(pca_data_transform)
lower_tri_idx <- lower.tri(corr_matrix)

comparison_table <- data.frame(
  Pair = paste(rownames(corr_matrix)[row(corr_matrix)[lower_tri_idx]], 
              "vs",
              colnames(corr_matrix)[col(corr_matrix)[lower_tri_idx]]),
  Original = corr_matrix[lower_tri_idx],
```

Table 1: Correlation Comparison (Pre vs. Post Transformation)

Pair	Original	Transformed	Change
deprawsc vs flourish	-0.508	-0.534	-0.025
anx_score vs flourish	-0.382	-0.416	-0.035
lonesc vs flourish	-0.467	-0.504	-0.036
ed_sde vs flourish	-0.172	-0.189	-0.016
anx_score vs deprawsc	0.748	0.759	0.012
lonesc vs deprawsc	0.540	0.550	0.010
ed_sde vs deprawsc	0.348	0.365	0.018
lonesc vs anx_score	0.475	0.487	0.013
ed_sde vs anx_score	0.323	0.340	0.018
ed_sde vs lonesc	0.265	0.265	0.000

```

    Transformed = corr_matrix_2[lower_tri_idx]
}

comparison_table$Change <- comparison_table$Transformed - comparison_table$Original

kable(comparison_table,
      digits = 3,
      format = "latex",
      booktabs = TRUE,
      caption = "Correlation Comparison (Pre vs. Post Transformation)")

```

Evidently, there is a

```
corrplot.mixed(corr_matrix_2, lower = "number", upper = "ellipse")
```

