Homework 3: Discriminant Analysis

S&DS 5360 | Multivariate Statistics

Brandon Tran (brandon.tran@yale.edu)

February 19, 2026

Setup & Data Overview

We continue with our analysis of the 2024–2025 Healthy Minds Study (HMS) student survey, this time performing discriminant analysis. We will investigate whether we can use the same psychological health scores we defined in Homework 2 to determine whether a student belongs to one of the following GSM identities:

1. Cis-Heterosexual
2. Cis-LGBQ+ (lesbian, gay, bisexual, or queer, but not TGNC)
3. Transgender and Gender Nonconforming (TGNC), including those who also identify as LGBQ+

Per the grader's feedback, we omit the code necessary to load and preprocess the data. However, we provide summary Table 1 contextualize the subsequent analyses:

Table 1: Descriptive Statistics by Identity Group

Statistic	Cis-Het	Cis-LGBQ+	TGNC
Sample Size			
n	44,791	16,387	5,155
Depression (PHQ-9)			
Depression Mean	7.26	9.93	12.06
Depression SD	5.80	6.16	6.42
Anxiety (GAD-7)			
Anxiety Mean	6.66	9.04	10.29
Anxiety SD	5.54	5.71	5.66
Loneliness (UCLA)			
Loneliness Mean	5.24	5.96	6.53
Loneliness SD	1.91	1.88	1.81
Flourishing			
Flourishing Mean	44.84	41.97	39.08
Flourishing SD	8.76	8.53	8.90

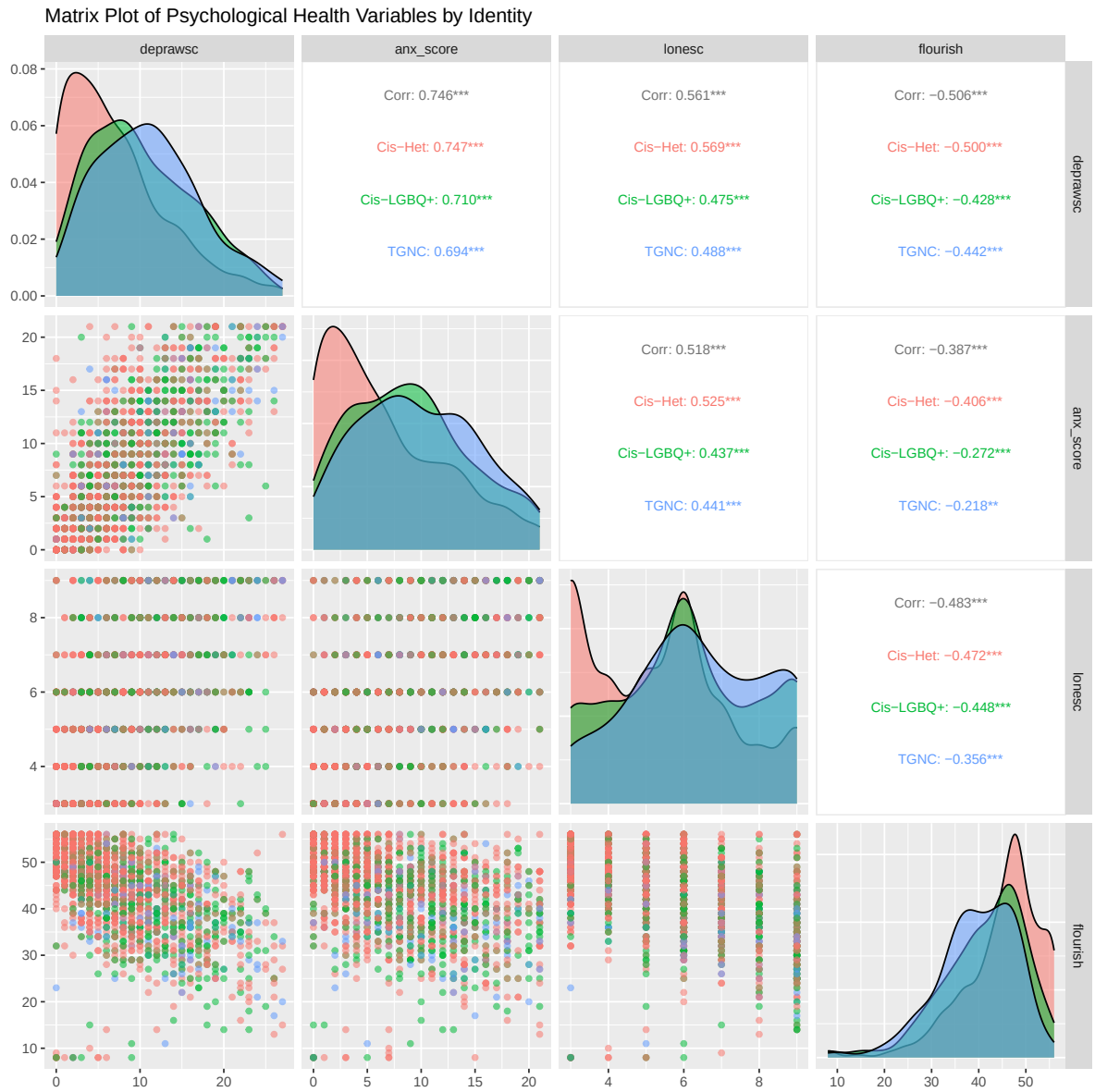
We visually observe a difference across psychological health score means for each group and note the significant sample size. We feel comfortable proceeding with our analysis.

Part 1 | Evaluating Assumptions

We begin by visualizing our data via a matrix (pairs) plot:

```
set.seed(4747)
df_sample <- df %>% sample_n(2000)

ggpairs(
    df_sample,
    columns = c('deprawsc', 'anx_score', 'lonesc', 'flourish'),
    aes(color = identity_group, alpha = 0.5),
    upper = list(continuous = wrap('cor', size = 3)),
    title = 'Matrix Plot of Psychological Health Variables by Identity'
)
```

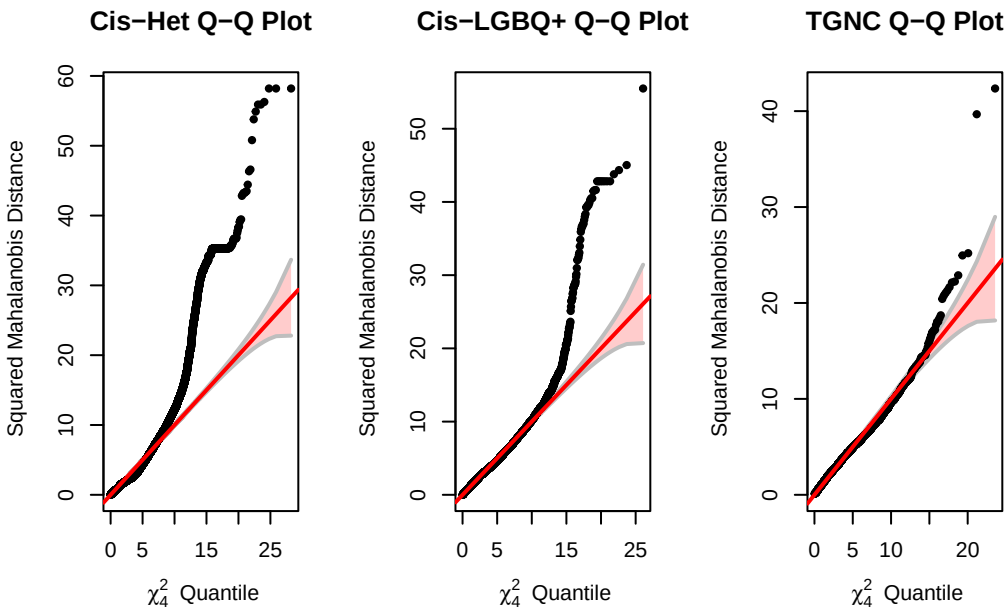


We see strong **separaton of means** for flourishing (`flourish`) and depression (`deprawsc`). In comparison, loneliness (`lonesc`) features a lot of overlap and likely will not be a strong predictor in our model.

Regarding **normality,** we first note a slight right skew for depression (`deprawsc`) and anxiety (`anx_score`), particularly for the Cis-Het group. The other groups appear more symmetrical. Loneliness (`lonesc`) is visibly multimodal, violating normality (likely due to the LIkert scale structure of the question), and potentially reducing the accuracy of the discriminant function. Finally, flourishing (`flourish`) is left-skewed. We formally analyze multivariate normality via

chi-square quantile plots.

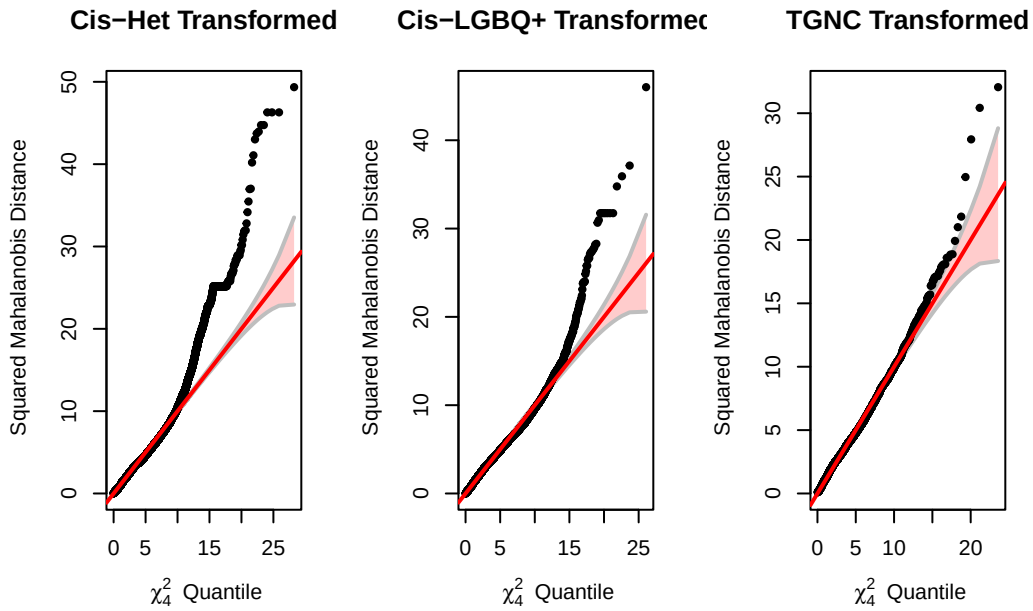
```
par(mfrow=c(1,3))
cqplot(df[df$identity_group == "Cis-Het",
    c("deprawsc", "anx_score", "lonesc", "flourish")], main = "Cis-Het Q-Q Plot")
cqplot(df[df$identity_group == "Cis-LGBQ+",
    c("deprawsc", "anx_score", "lonesc", "flourish")], main = "Cis-LGBQ+ Q-Q Plot")
cqplot(df[df$identity_group == "TGNC",
    c("deprawsc", "anx_score", "lonesc", "flourish")], main = "TGNC Q-Q Plot")
```



Our Chi-Square Quantile plots show heavy deviations from the diagonal in the upper tails for all groups, though less so for `TGNC`. This indicates that our data is not multivariate normal, likely due to the skew caused by Likert scale rankings. We consider the same transformations as in Homework 2:

- Square root of depression
- Square root of anxiety
- Square (quadratic) of flourish

However, these and similar transformations do not yield improved multivariate normality. We suppress the code for the ease of the grader but present the corresponding Q-Q plots for the transformed data:



While the Chi-Square Quantile plots indicate a violation of multivariate normality, we recall that discriminant analysis is traditionally robust to such violations in large samples like ours. The non-normality is an inherent feature of the floor effects of the Likert scale ratings, so while transformations were considered, we ultimately choose to proceed with the raw variables to preserve the interpretability of the results. The violation suggests that QDA could be tested as a potentially more flexible alternative to LDA, but ultimately, we feel comfortable proceeding with LDA.

Finally, we check for **equal covariance**. The matrix plot shows that anxiety versus depression maintains fairly equal correlation across all groups (0.747, 0.710, 0.694). This trend is reasonably consistent among other predictor pairs; however, flourishing versus depression, for example, features wider gaps in correlation $(0.472, 0.448, 0.356)$. These still feel reasonably equivalent, but for completeness, we formally test for equal covariance via Box's M-test.

```
boxM(df[, c("deprawsc", "anx_score", "lonesc", "flourish")], df$identity_group)
```

```
 Box's M-test for Homogeneity of Covariance Matrices

data:  df[, c("deprawsc", "anx_score", "lonesc", "flourish")] by df$identity_group
Chi-Sq (approx.) = 1198.817, df = 20, p-value = < 2.2e-16
```

As expected, given our large sample size of 66,333 (aggregate count by group can be seen in Table 1), leads our Box's M test to reject the null hypothesis, indicating a difference in covariances across groups. We instead analyze the practical difference in covariance structures.

```r
vars <- c("deprawsc", "anx_score", "lonesc", "flourish")

get_ratio <- function(m1, m2) {
  r <- m1 / m2
  r[abs(r) < 1] <- 1 / r[abs(r) < 1]
  return(round(r, 1))
}

# Calculate Covariance Matrices
cov_cis  <- cov(df[df$identity_group == "Cis-Het", vars])
cov_lgbq <- cov(df[df$identity_group == "Cis-LGBQ+", vars])
cov_tgnc <- cov(df[df$identity_group == "TGNC", vars])

# Calculate Ratios
rat_tgnc_cis  <- get_ratio(cov_tgnc, cov_cis)
rat_lgbq_cis  <- get_ratio(cov_lgbq, cov_cis)
rat_tgnc_lgbq <- get_ratio(cov_tgnc, cov_lgbq)
```

Table 2: Ratio of Covariance Matrix Elements Between Groups

Metric	Depression	Anxiety	Loneliness	Flourishing
Comparison 1: TGNC vs. Cis-Het				
Depression	1.2	1.0	1.1	1.4
Anxiety	1.0	1.0	1.3	1.1
Loneliness	1.1	1.3	1.1	1.1
Flourishing	1.4	1.1	1.1	1.0
Comparison 2: Cis-LGBQ+ vs. Cis-Het				
Depression	1.1	1.0	1.1	1.2
Anxiety	1.0	1.1	1.2	1.0
Loneliness	1.1	1.2	1.0	1.0
Flourishing	1.2	1.0	1.0	1.1
Comparison 3: TGNC vs. Cis-LGBQ+				
Depression	1.1	1.0	1.1	1.2
Anxiety	1.0	1.0	1.1	1.1
Loneliness	1.1	1.1	1.1	1.1
Flourishing	1.2	1.1	1.1	1.1

Following the rule where if all ratios of covariance matrix elements between groups are less

than four, the covariance matrices are similar enough, we feel comfortable proceeding with linear discrimant analysis, as all of our ratios fall between 1.0 and 1.4.

Part 2 | Discriminant Analysis

With our assumptions somewhat satisfied in Part 1, we proceed with both Linear Discriminant Analysis and Quadratic Discriminant Analysis. We will compare their classification accuracy and proceed with analysis via the outperforming method.

```
lda_cv <- lda(
    identity_group ~ deprawsc + anx_score + lonesc + flourish,
    data = df,
    prior = c(1/3, 1/3, 1/3),
    CV = TRUE)

qda_cv <- qda(
    identity_group ~ deprawsc + anx_score + lonesc + flourish,
    data = df,
    prior = c(1/3, 1/3, 1/3),
    CV = TRUE)

lda_cv_acc <- round(sum(
  diag(prop.table(table(df$identity_group, lda_cv$class)))), 4)
qda_cv_acc <- round(sum(
  diag(prop.table(table(df$identity_group, qda_cv$class)))), 4)

cat("LDA CV Accuracy:", lda_cv_acc, "\nQDA CV Accuracy:", qda_cv_acc)
```

```
LDA CV Accuracy: 0.5167
QDA CV Accuracy: 0.52
```

We see that the cross-validated accuracy of QDA is only 0.0033 higher than that of LDA. Therefore, to preserve interpretability (as discussed in Part 1) and to adhere with the principle of parsimony, we proceed with LDA on our raw (untransformed) variables.

Note that we compared LDA versus QDA as full models (inclusive of all discriminants). Now that we have elected to use LDA, we proceed with Stepwise Discriminant Analysis to determine whether there are any redundant discriminators we should drop. Note we proceed with 10-fold validation due to our large sample size.

```r
df_clean <- df %>% drop_na()

X <- df_clean[, c('deprawsc', 'anx_score', 'lonesc', 'flourish')]
y <- df_clean$identity_group

step_model <- stepclass(
    X,
    y,
    method = "lda",
    direction = "both",
    prior = c(1/3, 1/3, 1/3),
    improvement = 0.01,
    fold = 10,
    output = FALSE)

step_model
```

```
method      : lda
final model : y ~ flourish
<environment: 0x00000224f4565f20>

correctness rate = 0.517
```

With equal priors and a low improvement threshold, our Stepwise Discriminant Analysis proceeds with solely `deprawsc` as the discriminator, yielding essentially the same correctness rate as our full model (0.5172 versus 0.5167, respectively). This is unsurprising, as we saw that the discriminators themselves exhibit multicollinearity. Thus, one discriminator alone may already provide a bulk of the signal used to classify each individual. However, per the recommendation in the assignment text, we proceed with the full model (all four discriminators). This will allow us to conduct a more granular deep dive into the multivariate psychological health profiles of each group in later analyses. The model summary is produced below:

```r
lda_fit <- lda(identity_group ~ ., data = df, prior = c(1/3, 1/3, 1/3)); lda_fit
```

```
Call:
lda(identity_group ~ ., data = df, prior = c(1/3, 1/3, 1/3))

Prior probabilities of groups:
  Cis-Het Cis-LGBQ+      TGNC
0.3333333 0.3333333 0.3333333
```

```
Group means:
          deprawsc anx_score   lonesc flourish
Cis-Het    7.25590  6.662142 5.244625 44.84475
Cis-LGBQ+  9.93086  9.040459 5.959968 41.97223
TGNC      12.06324 10.293889 6.530747 39.08380

Coefficients of linear discriminants:
                  LD1          LD2
deprawsc  -0.08766862 -0.119020335
anx_score -0.01848018  0.238315655
lonesc    -0.16241115  0.009194239
flourish   0.03460451  0.054181217

Proportion of trace:
   LD1    LD2
0.9828 0.0172
```

Part 3 | Multivariate Group Means

To determine whether the multivariate group means are different, we perform the multivariate Wilks' Lambda Test:

```r
df_manova <- manova(as.matrix(df[, vars]) ~ df$identity_group)
summary(df_manova, test = "Wilks")
```

```
                     Df  Wilks approx F num Df den Df    Pr(>F)
df$identity_group     2 0.9191   714.34      8 132654 < 2.2e-16 ***
Residuals         66330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our Wilks' Lambda is 0.9191, indicating that approximately 8.1% of the total variance in the combined mental health scores is explained solely by the student's identity group. With an extremely small p-value, we can reject the null hypothesis and conclude that mean vectors are not the same across groups.

We further present the univariate analyses (equivalent to `summary.aov` output) to better understand the relevance of each discriminator in Table 3.

We see that each predictor independently is a statistically significant discriminator of group differences. Given our sample size and earlier discussion on collinearity, this is unsurprising.

Table 3: MANOVA and Univariate ANOVA Results for Identity Groups

Analysis Level	Wilks' Lambda	F-Statistic	p-value
Multivariate Test			
Multivariate (Wilks)	0.9191	714.34	$< 2.2e\text{-}16$
Univariate Response Tests			
Depression (PHQ-9)	—	2342.80	$< 2.2e\text{-}16$
Anxiety (GAD-7)	—	1766.50	$< 2.2e\text{-}16$
Loneliness (UCLA)	—	1645.50	$< 2.2e\text{-}16$
Flourishing	—	1432.40	$< 2.2e\text{-}16$

However, note that the F- statistic for depression is much larger than the other discriminators. This confirms the stepwise decision earlier to retain only depression in the model.

Ultimately, we are satisfied with our Wilks' Lambda value of 0.9191. Explaining 8.1% of the total variance across groups is generally considered a medium effect in the social sciences.

Part 4 | Signifiance of Discriminant Functions

Recall the proportion of trace metrics we computed in Part 2:

```
lda_fit
```

```
Call:
lda(identity_group ~ ., data = df, prior = c(1/3, 1/3, 1/3))

Prior probabilities of groups:
  Cis-Het Cis-LGBQ+      TGNC
0.3333333 0.3333333 0.3333333

Group means:
          deprawsc anx_score   lonesc flourish
Cis-Het    7.25590  6.662142 5.244625 44.84475
Cis-LGBQ+  9.93086  9.040459 5.959968 41.97223
TGNC      12.06324 10.293889 6.530747 39.08380

Coefficients of linear discriminants:
                  LD1          LD2
deprawsc  -0.08766862 -0.119020335
anx_score -0.01848018  0.238315655
```

```
lonesc    -0.16241115  0.009194239
flourish   0.03460451  0.054181217

Proportion of trace:
   LD1    LD2
0.9828 0.0172
```

Since we have 3 groups, we have 2 discriminant functions. Our first discriminant function (`LD1`) accounts for 98.28% of the relative discriminating power—almost all of the necessary discriminatory signal in our predictor variables. `LD2`, on the other hand, captures only the remaining 1.72%, though it is statistically significant (likely due to sample size). This suggests that the psychological differences between each group are largely one-dimensional.

Part 5 | Classification

Recall that we performed both regular and 10-fold cross validation to determine whether to proceed with LDA or QDA in Part 2.

```r
lda_regular_acc <- round(
  sum(diag(prop.table(table(df$identity_group, predict(lda_fit)$class)))), 4)

acc_comp <- data.frame(
    Type = c("Regular Accuracy", "Cross-Validated Accuracy"),
    Value = c(lda_regular_acc, lda_cv_acc)
)

acc_comp
```

```
                      Type  Value
1         Regular Accuracy 0.5168
2 Cross-Validated Accuracy 0.5167
```

Our regular accuracy and 10-fold cross-validated accuracy are nearly identical, with a difference of 0.0001. This indicates that our model is stable. Now, we are curious whether there are certain areas where the model does particularly well (or not). For this, we generate a confusion matrix and append recall (accuracy percentages):

```r
conf_matrix <- table(Actual = df$identity_group, Predicted = lda_cv$class)
```

Table 4: Cross-Validated Confusion Matrix for Identity Groups

	Predicted Group Membership			
Actual Group	Pred: Cis-Het	Pred: Cis-LGBQ+	Pred: TGNC	Recall (%)
Cis-Het	28363	6331	44791	63.32
Cis-LGBQ+	7040	3106	16387	18.95
TGNC	1498	852	5155	54.41

Though our overall cross-validated accuracy is $\approx 52\%$, as illustrated in Table 4 the model is most effective at predicting `Cis-Het` and `TGNC`, with recalls of 63.3% and 54.4%, accordingly. However, the `Cis-LGBQ+` group has a low recall of 19.0%. Most of these students are misclassified as `TGNC`, illustrating that their psychological health data doesn't form a unique cluster. Rather, they may exhibit scores that are more similar to the `TGNC` community, which is unsurprising, given both groups undergo systemic forms of inequity tied under the same "LGBTQIA+" umbrella.

Part 6 | "Best" Discriminators

We cannot rely solely on raw coefficients from `lda_fit` to determine the "best" discriminator(s), as our variables are scaled differently. We thus calculate standardized discriminant coefficients.

```
lda_std <- lda(identity_group ~
           scale(deprawsc) + scale(anx_score) + scale(lonesc) + scale(flourish),
           data = df_clean,
           prior = c(1/3, 1/3, 1/3))

std_coefs <- lda_std$scaling
```

Table 5: Standardized Discriminant Coefficients

	LD1 (98.3%)	LD2 (1.7%)
scale(deprawsc)	-0.539	-0.732
scale(lonesc)	-0.316	0.018
scale(flourish)	0.308	0.482
scale(anx_score)	-0.106	1.368

We see in Table 5 that depression (`deprawsc`) is the dominant predictor on the first (and second) discriminant function, with a coefficient of -0.539 after all variables are scaled. This

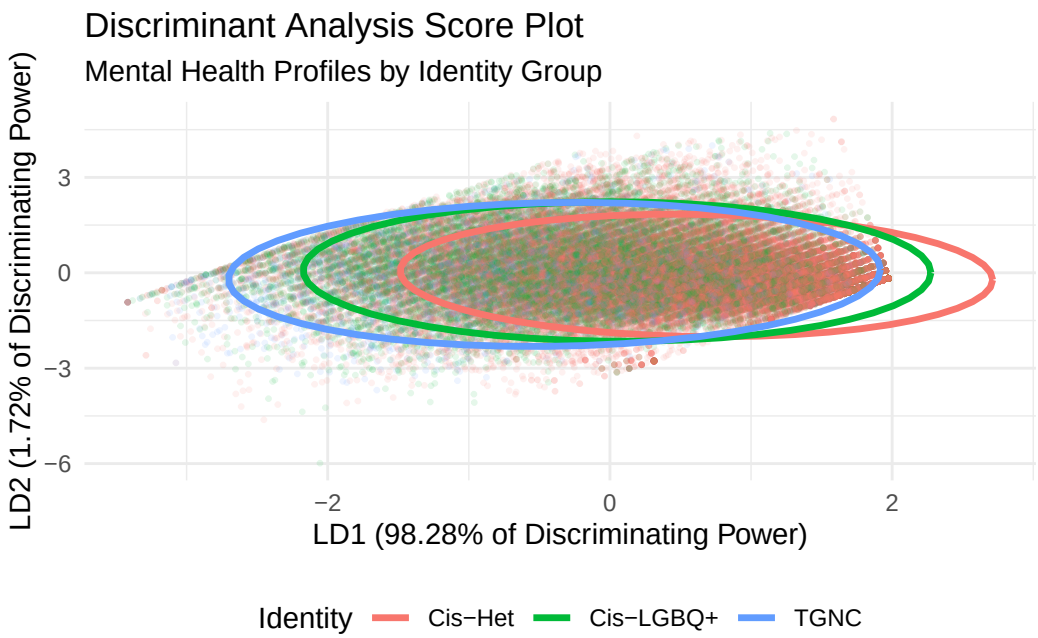
indicates that differences in depression are the primary metric separating `Cis-Het`, `Cis-LGBQ+`, and `TGNC` students. Loneliness and flourishing were the second most dominant predictors, with anxiety providing the least unique discriminatory power once the other variables are accounted for.

Part 7 | Score Plots

To visualize the results of our preceding analyses, we generate score plots. Note that to enhance visual interpretability, we use `ggplot2` as opposed to the plotting functons of base `R` that were used in class.

```
lda_preds <- predict(lda_fit)
score_data <- data.frame(
  Identity = df_clean$identity_group,
  LD1 = lda_preds$x[, 1],
  LD2 = lda_preds$x[, 2]
)

ggplot(score_data, aes(x = LD1, y = LD2, color = Identity)) +
  geom_point(alpha = 0.1, size = 0.5) +
  stat_ellipse(linewidth = 1.2, level = 0.95) +
  theme_minimal() +
  labs(
    title = "Discriminant Analysis Score Plot",
    subtitle = "Mental Health Profiles by Identity Group",
    x = "LD1 (98.28% of Discriminating Power)",
    y = "LD2 (1.72% of Discriminating Power)"
  ) +
  theme(legend.position = "bottom")
```



Our first observation is the horizontal dominance of `LD1`; that is, almost all of the separation between ellipses happens horizontally along the x-axis. This reinforces our finding in Part 4 that `LD1` accounts for a majority (98.28%) of our discriminating power. Conversely, `LD2` accounts for only 1.72%, evidenced by the lack of vertical separation of the ellipses along the y-axis.

Furthermore, recall our confusion matrix in Part 5. There, we observeed that our model frequently misclassifies Cis-LGBQ+ students, most often as TGNC. This reflects in our score plot, which reveals that this group exhibits psychological health scores that exist in between Cis-Het and TGNC students, with more overlap with the latter as opposed to the former. We therefore feel comfortable with the results we have encountered thus far.

Part 8 | Visualizing Discriminants

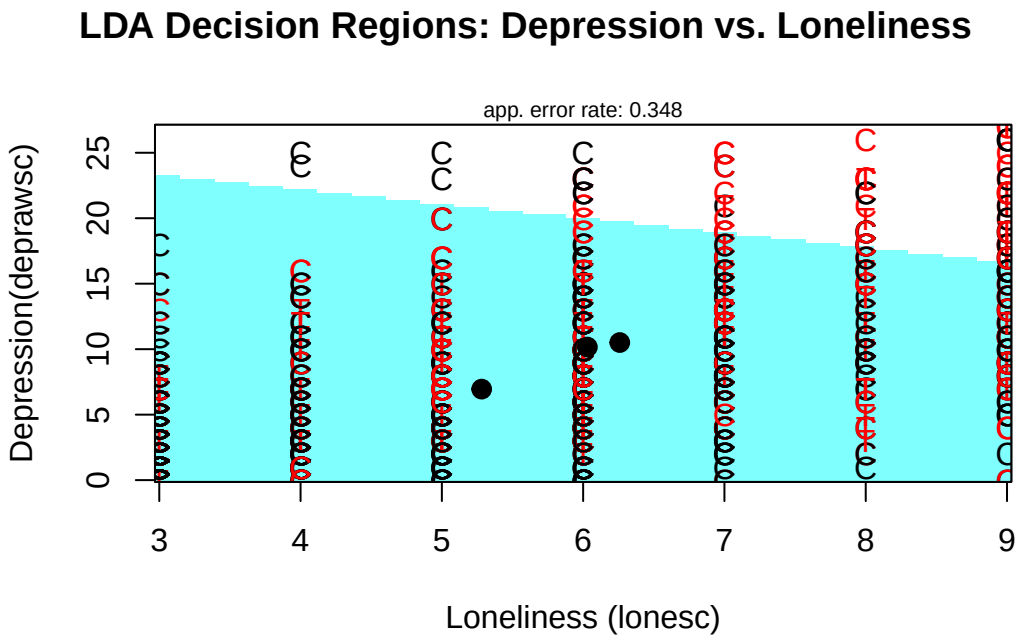
We will use the `partimat()` function to plot the shaded decision regions for a pair of our original variables: specifically, of depression (`deprawsc`) and loneliness (`lonesc`), which were identified as our "best" discriminators in Part 6. Note that we will need to take a sample of our data, as we will suffer from overplotting and rendering issues otherwise.

```
set.seed(4747)

df_sample <- df %>% sample_n(1000) %>%
```

```
    as.data.frame()

partimat(
    identity_group ~ deprawsc + lonesc,
    data = df_sample,
    method = 'lda',
    nplots.vert = 1,
    nplots.hor = 1,
    main = 'LDA Decision Regions: Depression vs. Loneliness',
    name = c('Depression(deprawsc)', 'Loneliness (lonesc)')
    )
```



The resulting partition plot displays the decision regions spanning our two strongest discriminators. It reveals a linear boundary separating the low- distress region (bottom) from the high distress region (top). Notably, while we have three identity groups, only two prediction territories emerge. The centroids align linearly, confirming the dominance of the first discriminant function (`LD1`). However, the lack of a distinct prediction region for the middle group (`Cis-LGBQ+`), as well as the high volume of misclassified points, visually reinforces our cross-validated accuracy results. While identity groups show significantly different mean mental health profiles, individual variance is too large to cleanly and linearly separate them using only these two metrics.

Part 9 | k-Nearest Neighbors

We will now attempt to perform classification via a popular nonparametric method, k-nearest neighbors. This provides us the flexibility of nonlinear decision boundaries and circumvents the need for equal covariances amongst discriminants. Note that to properly perform k-NN, we must scale our data and leverage cross-validation. We set $k = 50$ to match our sample size.

```
# Create an 80-20 train/test split.
set.seed(4747)
train_idx <- sample(1:nrow(df), size = 0.8 * nrow(df))

# Scale predictors.
train_x <- scale(df[train_idx,
                c('deprawsc', 'anx_score', 'lonesc', 'flourish')])
test_x <- scale(df[-train_idx,
                c('deprawsc', 'anx_score', 'lonesc', 'flourish')],
                center = attr(train_x, 'scaled:center'),
                scale = attr(train_x, 'scaled:scale'))

# Isolate response.
train_y <- df$identity_group[train_idx]
test_y <- df$identity_group[-train_idx]

# Run kNN.
knn_pred <- knn(train = train_x, test = test_x, cl = train_y, k = 50)

# Output results.
knn_ct <- table(Actual = test_y, Predicted = knn_pred)
knn_acc <- round(sum(diag(prop.table(knn_ct))), 4); knn_acc
```

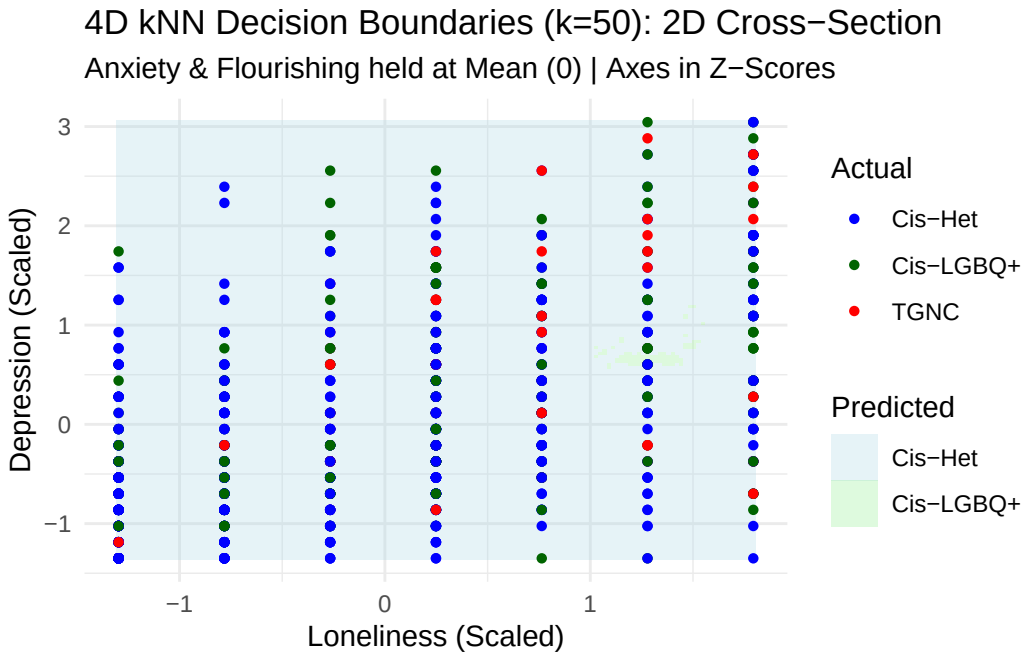
```
[1] 0.6768
```

Table 6: kNN Confusion Matrix for Identity Groups (Accuracy = 67.68%)

	Predicted Group Membership			
Actual Group	Pred: Cis-Het	Pred: Cis-LGBQ+	Pred: TGNC	Recall (%)
Cis-Het	8820	212	4	97.61
Cis-LGBQ+	3019	157	7	4.93
TGNC	953	93	2	0.19

We see a huge jump in predictive accuracy to 67.68%, which appears to be much higher than our accuracy rate yielded by LDA of 51.67%. However, when looking at the confusion matrix in Table 6, we see that this is due to dominance from `Cis-Het`, which happens to comprise 67.52% of our data. Said differently, kNN is just predicting `Cis-Het` over and over again, yielding a high recall for that group of 97.61% but low recalls for the other groups. We were able to control for this with LDA by setting equal priors, but kNN does not use such priors.

To compare against the decision regions we plotted for our strongest two discriminators in Part 8 (depression and loneliness), we plot the decision boundaries yielded by kNN for these same variables. We suppress the code to generate the plot for the grader's convenience but note that `ggplot` is used.



We see, as insinuated by the confusion matrix, a vast majority of the decision regions predict `Cis-Het`, resulting in 68% accuracy because 67% of our data is, in fact, `Cis-Het`. There is only one region (in green) where the model predicts `Cis-LGBQ+`, and `TGNC` appears to never be predicted.

This suggests that for highly overlapping, imbalanced demographic data, enforcing equal priors via discriminant analysis yields a much more practically useful and balanced classification model.