Assignment 2

Madimetja Maredi 2014095653

08 August 2025

Introduction: Aims and Purpose

This analysis investigates the relationship between problem behaviours and environmental factors in a university student population. Given the complexity of interpreting 14 variables simultaneously, Canonical Correlation Analysis (CCA) is used to identify key patterns of association between the two sets.

The aims are to:

- Extract pairs of canonical variates that capture the strongest links between the two variable sets.
- Interpret these variates through their correlations with the original variables.
- Summarize how specific psychosocial factors relate to patterns of problem behaviours.

This approach provides a structured understanding of how students' behavioural risks align with their personal and environmental profiles.

Materials and Methods

- **Data Preparation**

The dataset includes 498 students and contains no missing values. All variables were suitable for multivariate analysis.

- **Variable Grouping**

Variables were grouped into two sets:

Set 1: Problem Behaviours (6 variables)

CIAS_Total, Tobacco_Use, Alcohol_Use, Illicit_Drug_Use, Unprotected_Sex, Gambling_Behaviour

Set 2: Personal & Environmental Factors (8 variables)

Impulsivity, Social_Interaction_Anxiety, Depression, Social_Support, Intolerance_of_Deviance, Family_Morals, Family_Conflict, GPA

These sets reflect core behavioural and psychosocial dimensions relevant to student well-being.

The main R packages used were `CCA` for the core analysis, `CCP` for significance testing, `psych` for descriptive statistics, and `corrplot` for visualisation.

Preliminary Analysis

1. Data Structure and Descriptive Statistics

```
## Total missing values in the dataset: 0
```

```
## Number of observations (students): 498
## Number of variables in Set 1 (Problem Behaviours): 6
```

```
## Number of variables in Set 2 (Personal/Environmental Factors): 8
```

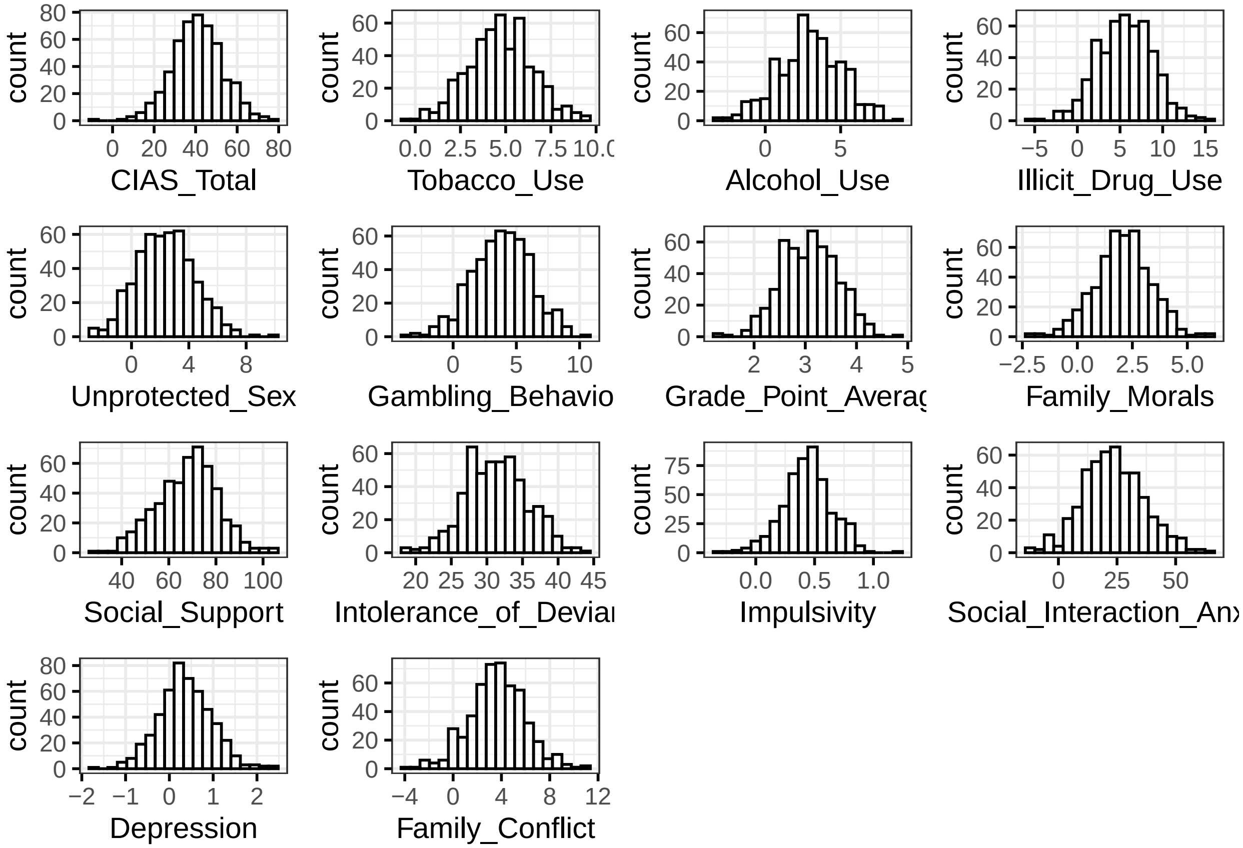
The dataset contains **498 observations and 14 variables**. There are **no missing values**. The variables are divided into a set of **6 problem behaviours** and a set of **8 environmental factors**.

Table 1: Summary of the Data

	n	mean	sd	min	max	skew
CIAS_Total	498	40.81	12.12	-9.09	77.52	-0.06
Tobacco_Use	498	4.71	1.78	-0.60	9.35	-0.02
Alcohol_Use	498	2.96	2.11	-3.42	8.50	-0.14
Illicit_Drug_Use	498	5.59	3.22	-5.44	15.63	-0.06
Unprotected_Sex	498	2.38	1.99	-2.94	9.61	0.13
Gambling_Behavior	498	3.82	2.30	-3.66	10.55	-0.14
Grade_Point_Average	498	3.07	0.55	1.25	4.76	-0.06
Family_Morals	498	2.06	1.29	-2.02	6.16	-0.10
Social_Support	498	67.90	12.80	28.66	104.97	-0.13
Intolerance_of_Deviance	498	31.30	4.43	18.53	43.78	0.02
Impulsivity	498	0.44	0.21	-0.29	1.24	-0.15
Social_Interaction_Anxiety	498	23.40	13.12	-13.85	62.84	0.15
Depression	498	0.36	0.61	-1.81	2.30	0.12
Family_Conflict	498	3.63	2.33	-3.80	11.08	-0.06

Table 1 above shows that the variables are measured on different scales. For instance, `Social_Support` has a mean of 67.90, while `Impulsivity` has a mean of 0.44. This confirms the necessity of using a method like CCA that standardizes the variables to ensure each contributes appropriately to the analysis.

2. Normality check



The output above displays the histograms of all 14 variables used in the analysis. Visual inspection shows that the majority of the variables are approximately normally distributed or only mildly skewed. While minor skewness may still be present, it does not appear severe enough to justify transformation. Therefore, based on the histogram evidence, we proceed with the original (untransformed) variables for Canonical Correlation Analysis without applying any transformations.

3. Within-Set Correlations

Table 2: Correlation Matrix - Environmental Factors

	GPA	Morals	Support	Deviance	Impulsivity	Anxiety	Depression	Conflict
GPA	1.00	-0.01	-0.09	0.17	-0.16	0.10	0.07	0.07
Morals	-0.01	1.00	0.12	0.09	0.02	-0.06	-0.03	-0.07
Support	-0.09	0.12	1.00	0.11	0.00	-0.34	-0.28	-0.27
Deviance	0.17	0.09	0.11	1.00	-0.32	0.09	0.07	-0.21
Impulsivity	-0.16	0.02	0.00	-0.32	1.00	-0.03	0.11	0.17
Anxiety	0.10	-0.06	-0.34	0.09	-0.03	1.00	0.29	0.22
Depression	0.07	-0.03	-0.28	0.07	0.11	0.29	1.00	0.16
Conflict	0.07	-0.07	-0.27	-0.21	0.17	0.22	0.16	1.00

Table 3: Correlation Matrix - Problem Behaviour Variables

	CIAS	Tobacco	Alcohol	Drugs	Unprotected	Gambling
CIAS	1.00	0.46	0.24	0.21	0.36	-0.09
Tobacco	0.46	1.00	0.37	0.13	0.46	-0.15
Alcohol	0.24	0.37	1.00	0.34	0.42	-0.01
Drugs	0.21	0.13	0.34	1.00	0.35	0.00
Unprotected	0.36	0.46	0.42	0.35	1.00	-0.11
Gambling	-0.09	-0.15	-0.01	0.00	-0.11	1.00

Table 2 and **Table 3** present the correlations within each of the two variable sets used in the Canonical Correlation Analysis.

Within **Set 1 (Problem Behaviours)**, some moderate relationships are observed. For example, `CIAS` shows moderate positive correlations with both `Tobacco use` (r = 0.46) and `Unprotected sex` (r = 0.36), while `Tobacco`, `Alcohol`, and `Unprotected sex` are also fairly interrelated. These patterns suggest that certain risky behaviours tend to cluster together. On the other hand, `Gambling` stands apart, showing weak or even negative associations with most other variables, indicating it may represent a different behavioural domain.

In **Set 2 (Environmental Factors)**, the relationships are generally weaker but still meaningful. For instance, `Anxiety` and `Depression` are positively related (r = 0.29), and both are negatively associated with `Support`, suggesting that lower perceived social support is linked with higher emotional distress. Conflict also shows small to moderate associations with several psychological factors, further reinforcing the idea of an underlying psychosocial dynamic.

Altogether, these within-set correlations show that the variables are reasonably related, which supports the decision to use Canonical Correlation Analysis to examine how these two sets of factors connect.

Major Analysis

1. Canonical Correlations

Table 4: Canonical Correlations

x
0.5746401
0.4851276
0.2834954
0.2117141
0.1943604
0.0261275

From the table, the highest canonical correlation (0.5746) is observed for the first pair of canonical variates, indicating that this pair captures the strongest linear relationship between the two variable sets. The second pair also shows a moderate correlation of 0.4851, while the remaining pairs display considerably lower values.

2. Significance of canonical correlations

To determine the number of canonical functions to retain, we use Wilks' Lambda to test their statistical significance. Only functions with **$p < 0.0001$** are considered statistically significant and retained for interpretation.

```
## Wilks' Lambda, using F-approximation (Rao's F):
```

```
##                stat   approx df1      df2      p.value
## 1 to 6:  0.4325916 9.227279  48 2385.545 0.000000e+00
## 2 to 6:  0.6458628 6.391480  35 2042.641 0.000000e+00
## 3 to 6:  0.8446501 3.505444  24 1696.661 2.296639e-08
## 4 to 6:  0.9184670 2.805080  15 1344.794 2.560957e-04
## 5 to 6:  0.9615672 2.414218   8  976.000 1.390712e-02
## 6 to 6:  0.9993174 0.111347   3  489.000 9.534598e-01
```

Examining the Wilks' Lambda output:

- Row 1 evaluates the null hypothesis that all six canonical correlations (Functions 1 to 6) are simultaneously zero. The test yields a highly significant result (p < 0.0001), indicating that at least one canonical function explains a meaningful portion of the shared variance. Thus, we reject the null hypothesis.
- Row 2 tests the significance of the remaining five canonical correlations (Functions 2 to 6) after removing the first. The result remains highly significant (p < 0.0001), so we reject the null hypothesis and retain the second function.
- Row 3 examines the remaining four correlations (Functions 3 to 6) after excluding the first two. The test yields a p-value of 2.30e-08, still below our threshold of 0.0001, indicating that the third canonical function is also statistically significant.
- Row 4 evaluates Functions 4 to 6, yielding a p-value of 0.000256, which exceeds the specified significance level. We therefore fail to reject the null hypothesis, suggesting that the fourth function is not statistically significant.
- Rows 5 and 6 further test Functions 5 to 6 and Function 6 individually. Both p-values (0.0139 and 0.9535, respectively) are well above the 0.0001 threshold, confirming that these functions are not significant.

Based on these results, we retain the first three canonical functions for interpretation.

3. Correlations of variables and canonical variables

Table 5: Correlation between Set 1 variables and canonical variables

	U1	U2	U3	U4	U5	U6
CIAS	0.74	-0.39	0.03	-0.42	0.26	0.24
Tobacco	0.63	-0.12	-0.40	0.43	0.50	0.05
Alcohol	0.70	0.06	0.39	0.52	-0.27	0.10
Drugs	0.55	0.01	-0.43	-0.09	-0.69	-0.16
Unprotected	0.71	0.10	0.07	0.02	0.17	-0.68
Gambling	0.07	0.89	-0.01	-0.27	0.02	0.36

Table 5 display the correlations between Set 1 variables and the canonical variates. The first canonical variate, U1, has strong positive correlations with CIAS (0.74), Alcohol use (0.70), Unprotected behaviour (0.71), and Tobacco use (0.63), indicating that higher scores on U1 correspond to greater involvement in these problem behaviours. The second canonical variate, U2, is most strongly associated with Gambling (0.89), suggesting this variate captures variation mainly related to gambling activity. Correlations with other variables on U2 and the later canonical variates are generally weaker or mixed.

Table 6: Correlation between Set 2 variables and canonical variables

	V1	V2	V3	V4	V5	V6
GPA	-0.40	0.24	0.14	-0.08	-0.72	-0.02
Morals	-0.20	-0.01	0.33	-0.38	0.06	0.14

	V1	V2	V3	V4	V5	V6
Support	0.02	-0.38	-0.22	-0.74	0.15	-0.44
Deviance	-0.82	-0.03	-0.46	-0.18	0.09	0.28
Impulsivity	0.70	0.14	-0.51	-0.07	-0.24	0.19
Anxiety	-0.24	0.77	-0.18	0.32	0.08	-0.40
Depression	0.08	0.76	0.08	-0.32	0.01	0.40
Conflict	0.21	0.45	0.04	0.24	0.37	0.20

Table 6 present the correlations between Set 2 variables and the canonical variates. The first canonical variate, V1, is strongly negatively correlated with Deviance (-0.82) and moderately negatively correlated with GPA (-0.40) and Morals (-0.20), while showing a strong positive correlation with Impulsivity (0.70). This suggests that higher V1 scores reflect higher impulsivity but lower deviance and academic performance. The second canonical variate, V2, is positively correlated with Anxiety (0.77) and Depression (0.76), indicating that V2 primarily captures internalizing emotional factors. Other variables show moderate or mixed correlations with subsequent canonical variates.

4. Variance shared between canonical variables

Table 7: Variance Shared Between Canonical Variables (Squared Canonical Correlations)

Canonical_Function	Rc2
Function 1	0.330
Function 2	0.235
Function 3	0.080
Function 4	0.045
Function 5	0.038
Function 6	0.001

Table 7 shows the variance shared between the canonical variates for each function. For the first canonical function, U1 and V1 share 33.0% of their variance. Function 2 accounts for 23.5%, while the remaining functions explain progressively less.

These results suggest that only the first two canonical functions reflect meaningful relationships between the two sets of variables. Functions 3 to 6 contribute little shared variance and are not retained for further interpretation.

5. Variance shared between canonical variables and their own variables

Table 8: Variance Shared Between Canonical Variables and Their Own Variable Sets

Canonical_Function	Prop_Var_Set1	Prop_Var_Set2
Function 1	0.374	0.183
Function 2	0.162	0.201
Function 3	0.083	0.086
Function 4	0.118	0.125
Function 5	0.149	0.095
Function 6	0.113	0.087

Table 8 shows how much variance each canonical variate shares with its own variable set. For example, U1 explains 37.4% of the total variance in Set 1 (problem behaviours), while V1 explains 18.3% of the variance in Set 2 (environmental factors).

This means U1 is a strong summary of its original set, while V1 provides a moderate summary.

Function 2 still represents a meaningful amount of variance in both sets (16.2% and 20.1%). However, the remaining functions (3 to 6) show low proportions of variance explained (under 15%), suggesting weak representation of their own variable sets.

Additional Analysis

1. Canonical Variable Coefficients & Equations

Table 9: Canonical Coefficients for Set 1 (Problem Behaviours)

CIAS	0.253	-0.255
Tobacco	0.082	0.029
Alcohol	0.107	0.008
Drugs	0.097	-0.014
Unprotected	0.130	0.167
Gambling	0.014	0.074

Table 10: Canonical Coefficients for Set 2 (Environmental Factors)

GPA	-0.389	0.289
Morals	-0.128	0.043
Support	0.007	0.003
Deviance	-0.139	-0.024
Impulsivity	2.164	0.252
Anxiety	-0.014	0.044
Depression	0.250	0.928
Conflict	0.019	0.086

Using **Tables 5 and 6**, the equations for the first canonical pair are:

- **U1** = 0.253* `CIAS_Total` + 0.082 * `Tobacco_Use` + …
- **V1** = -0.389 * `Grade_Point_Average` + 2.164 * `Impulsivity` + …

2. Canonical Variable Scores

```
##           U1          U2         U3         U4          U5          U6
## 1 -0.1580028 -0.29634932 -0.2448852  0.4022553 -1.08316882 -0.39433580
## 2 -1.4764600 -0.02486348  2.1366680 -2.4897068 -0.36148831 -0.40707794
## 3 -0.8339044 -0.44825514  1.0929796 -0.1210986 -1.07855634 -0.06103567
## 4 -1.4169712  0.01390068 -0.1360027 -0.3150809  0.08230914  0.78484574
## 5  0.3946926 -1.74526066 -1.4291919 -1.8532850 -1.29718769 -0.28202726
## 6 -1.2477768  1.60347677 -1.1097824  0.2274081 -0.25230992  0.86966400
##           V1         V2          V3          V4         V5         V6
## 1 -0.5632636 -0.2933189  0.74485188 -0.86926986 -1.8159878 -1.3222842
## 2 -0.9795355  0.4590528 -0.04731679  0.01343038 -1.5398894  1.2145312
## 3 -2.7157156  0.9533108 -0.89857837 -1.19406618  0.4169230 -0.9492970
## 4 -0.8613820 -1.0122534 -0.04166653 -0.20426301  0.1272574 -1.5892843
## 5 -0.4547013 -0.4318949 -0.11970172  1.28574388  0.3565625  0.5849706
## 6 -1.3926395  0.7842860  0.50418423  0.96444158  0.2942300 -0.6112948
```

The output shows the canonical variable scores. Looking at the first canonical variable pair, most students have negative scores on U1 and V1, indicating generally lower problem behaviours and associated risk factors. For the second canonical variable pair, scores vary widely on U2 and V2, reflecting individual differences on secondary behavioural and environmental dimensions. Scores on subsequent canonical variate pairs (U3 to U6 and V3 to V6) show mixed positive and negative values, capturing more subtle variations across students.

Conclusion

This Canonical Correlation Analysis revealed that problem behaviours and psychosocial factors among university students are interconnected along at least two major dimensions:

- The first dimension links higher impulsivity, lower deviance intolerance, and lower academic performance with greater engagement in online addiction, substance use, and risky sexual behaviour.
- The second dimension is shaped mainly by gambling behaviour, associated with internalising symptoms such as anxiety and depression.

Together, the first two canonical functions explain the most meaningful shared variance between the variable sets, while also representing substantial proportions of variance within their own sets. These findings point to distinct psychosocial profiles underpinning different types of problem behaviours.

Importantly, this analysis simplifies a complex web of 14 variables into just a few interpretable patterns, making it easier to understand the types of students who may be at greater risk and the nature of their underlying challenges. Such insights can be useful for targeted interventions in university health and counselling services.

R Code

```r
# --- Load libraries ---
library(CCA)
library(CCP)
library(psych)
library(knitr)
library(kableExtra)
library(corrplot)
library(ggpubr)
library(gridExtra)

# --- Load and prepare data ---
data <- read.csv("PIU.txt")
data <- data[ , -1]  # remove ID column if present

# Define the two variable sets exactly as used in the assignment
prob_behav <- data[, c("Tobacco_Use", "Alcohol_Use", "Illicit_Drug_Use",
                       "Gambling_Behavior", "Unprotected_Sex", "CIAS_Total")]
pers_env   <- data[, c("Grade_Point_Average", "Family_Morals", "Social_Support",
                       "Intolerance_of_Deviance", "Impulsivity",
                       "Social_Interaction_Anxiety", "Depression", "Family_Conflict")]

# --- Preliminary analysis ---

cat("Total missing values:", sum(is.na(data)), "\n")
cat("Number of observations:", nrow(data), "\n")
cat("Variables in Problem Behaviours set:", ncol(prob_behav), "\n")
cat("Variables in Personal/Environmental set:", ncol(pers_env), "\n")

# Descriptive statistics
descriptives <- psych::describe(data)
kable(descriptives[, c("n", "mean", "sd", "min", "max", "skew")],
      caption = "Table 1: Summary of the Data", digits = 2) %>% kable_styling(full_width = FALSE)

# Histograms (visual normality check)
plot_list <- lapply(names(data), function(var_name) {
  gghistogram(data, x = var_name, ggtheme = theme_bw(), bins = 20)
})
# Arrange histograms in grid (adjust ncol/nrow if needed)
ggarrange(plotlist = plot_list, ncol = 4, nrow = 4)

# Within-set correlation matrices
colnames(prob_behav) <- c("CIAS", "Tobacco", "Alcohol", "Drugs", "Unprotected", "Gambling")
colnames(pers_env)   <- c("GPA", "Morals", "Support", "Deviance", "Impulsivity", "Anxiety", "Depression

cor_prob <- round(cor(prob_behav), 2)
cor_pers <- round(cor(pers_env), 2)

kable(cor_prob, caption = "Table 2: Correlation Matrix - Problem Behaviour Variables", digits = 2) %>% k
kable(cor_pers, caption = "Table 3: Correlation Matrix - Personal & Environmental Factors", digits = 2)

# Corrplots side-by-side
par(mfrow = c(1, 2), mar = c(1,1,4,1))
corrplot(cor_prob,  method = "pie", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, title = "\nA) Problem Behaviours")
corrplot(cor_pers,  method = "pie", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, title = "\nB) Personal & Environmental Factors")
par(mfrow = c(1,1))

# --- Major analysis: Canonical Correlation Analysis ---
# Run CCA (note: cc() expects raw matrices; scaling optional)
cc_results <- cc(prob_behav, pers_env)

# Canonical correlations
cor_data <- data.frame(Canonical_Correlation = cc_results$cor)
kable(cor_data, caption = "Table 4: Canonical Correlations", digits = 4) %>% kable_styling(full_width =

# Significance testing (Wilks' Lambda; Rao's F approximation)
p <- ncol(prob_behav)
q <- ncol(pers_env)
n <- nrow(data)
wilks_results <- p.asym(cc_results$cor, n, p, q, tstat = "Wilks")
# wilks_results is a list; coerce relevant part to data.frame for printing
wilks_df <- as.data.frame(wilks_results[1:6])
rownames(wilks_df) <- paste0("Functions ", c("1-6","2-6","3-6","4-6","5-6","6-6"))
kable(wilks_df, caption = "Table 5: Wilks' Lambda Significance Tests (Rao's F)", digits = 6) %>% kable_
```

```
cc_summary <- data.frame(
  Canonical_Function = paste0("Function ", 1:length(cc_results$cor)),
  Canonical_Correlation = round(cc_results$cor, 4),
  Rc2 = round(cc_results$cor^2, 3)
)
kable(cc_summary, caption = "Table 6: Canonical Correlations and Rc^2", digits = 4) %>% kable_styling(fu

# --- Correlations of original variables with canonical variates ---
Corr1 <- cc_results$scores$corr.X.xscores
colnames(Corr1) <- paste0("U", 1:ncol(Corr1))
kable(round(Corr1, 2), caption = "Table 7: Correlation between Set 1 variables and canonical variables"
corrplot(Corr1, method = "number", is.corr = TRUE, title = "Figure 1: Set 1 v Canonical Variates")

Corr2 <- cc_results$scores$corr.Y.yscores
colnames(Corr2) <- paste0("V", 1:ncol(Corr2))
kable(round(Corr2, 2), caption = "Table 8: Correlation between Set 2 variables and canonical variables"
corrplot(Corr2, method = "number", is.corr = TRUE, title = "Figure 2: Set 2 v Canonical Variates")

# --- Variance shared between canonical variables (Rc^2) ---
canonical_cor <- cc_results$cor
Rc2 <- canonical_cor^2
Rc2_table <- data.frame(
  Canonical_Function = paste0("Function ", 1:length(Rc2)),
  Rc2 = round(Rc2, 3)
)
kable(Rc2_table, caption = "Table 9: Variance Shared Between Canonical Variables (Rc^2)", digits = 3) %

# --- Variance shared between canonical variables and their own variables (redundancy) ---
x_loadings <- Corr1   # already corr.X.xscores
y_loadings <- Corr2   # already corr.Y.yscores

prop_var_x <- colMeans(x_loadings^2)
prop_var_y <- colMeans(y_loadings^2)

red_table <- data.frame(
  Canonical_Function = paste0("Function ", 1:length(Rc2)),
  Rc2 = round(Rc2, 3),
  Prop_Var_Set1 = round(prop_var_x, 3),
  Redundancy_Set1 = round(prop_var_x * Rc2, 3),
  Prop_Var_Set2 = round(prop_var_y, 3),
  Redundancy_Set2 = round(prop_var_y * Rc2, 3)
)
kable(red_table, caption = "Table 10: Variance and Redundancy Indices", digits = 3) %>% kable_styling(fu

# --- Coefficients and equations ---
x_weights <- cc_results$xcoef[, 1:3]  # show first three if available
y_weights <- cc_results$ycoef[, 1:3]
kable(round(x_weights, 3), caption = "Table 11: Canonical Coefficients for Problem Behaviours (first 3
kable(round(y_weights, 3), caption = "Table 12: Canonical Coefficients for Personal/Environmental Facto


# U1 = sum( x_weights[,1] * corresponding original variables )
# V1 = sum( y_weights[,1] * corresponding original variables )
# --- Canonical scores ---
scores_all <- data.frame(cc_results$scores$xscores, cc_results$scores$yscores)
num_funcs <- ncol(cc_results$scores$xscores)
colnames(scores_all) <- c(paste0("U", 1:num_funcs), paste0("V", 1:num_funcs))
kable(head(scores_all), caption = "Table 13: Canonical Variable Scores (first 6 observations)", digits =

# Scatterplot of canonical variates for Function 1 (U1 vs V1)
U <- as.data.frame(cc_results$scores$xscores)
V <- as.data.frame(cc_results$scores$yscores)

ggscatter(data = data.frame(U1 = U$xscores.U1, V1 = V$yscores.V1),
          x = "U1", y = "V1", add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "U1 (Problem Behaviours)", ylab = "V1 (Environmental Factors)")


save(cc_results, scores_all, file = "CCA_results.RData")
write.csv(red_table, file = "redundancy_table.csv", row.names = FALSE)
```