1 The use of artificial intelligence (AI)

2 Load R-packages

3 Load the data

4 Introduction

5 ANOVA

6 Classification

7 Principal Component Analysis

8 References

Compulsory assignment, STAT340

Code ▾

Ane Kleiven

2025-03-28

1 The use of artificial intelligence (AI)

Chat-GPT has been used to help with some coding, especially solving problems with the code. It has also been useful when interpreting the results from the different analyses. Github-copilot was acitvated in R-studio, and has been used to help with some coding.

2 Load R-packages

For this assignment, several packages in R will be used. The packages are loaded in the code chunk below.

```r
library(kableExtra)
library(tidyverse)
library(ggplot2)
library(GGally)
library(BIAS.data)
library(mixlm)
library(MASS)
library(caret)
library(pls)
```

3 Load the data

The data used in this assignment is loaded in the code chunk below.

```r
load('data/CompulsorySTAT340.Rdata')
```

4 Introduction

In this report the data from the article “Does academia disfavor contextual and extraverted students?” (1) will be explored and analyzed using different statistical methods. The training set contains information about 5000 different students and 37 different variables. The test set is of equal size, and contains the same variables.

The variables are divided into 6 groups:

1. Demographic variables
2. Personality variables
3. Dichotomy pairs for personalities
4. Work type interests
5. Preferred instruction style
6. STEM subject interests

Table 1 shows the different variables in the data set, including long- and short names. The data will be further explained in the following sections, where the different statistical methods will be presented.

```
kable(Variables_names,
      caption = "Table 1: The different variables in the data set") |>
      kable_styling(bootstrap_options = "bordered", full_width = F, position = 'left')
```

Table 1: The different variables in the data set

	LongName	ShortName
1	Sex	Sex
2	Age	Age
4	Introvert	I
6	Sensing	S
8	Feeling	F
10	Contextual	C
12	Four Letter Type	type
13	Practical People Oriented Service	W1
14	Practical Technical Work	W2
15	Creative Enrepeneur Work	W3
17	Idealistiv Value Oriented Work	W5
18	Gross Motor Work	W6
20	Logical System Technical Work	W8
21	Communication and Informative Work	W9

	LongName	ShortName
23	Design Oriented Work	W11
24	Technical Computations	W12
26	Technical Problem Solving	W14
27	Development and Change Oriented Work	W15
28	Technical System Control	W16
29	Humanitarian Service Work	W17
31	Chaos and Change Oriented Work	W19
32	Mathematical Computations	W20
33	Practical Inductive Instruction	I1
35	Abductive Theoretical Instruction	I3
36	Dialog and Group Oriented Theoretical Instruction	I4
38	Methodological Formula Oriented Instruction	I6
39	Explorative Inspirational Instruction	I7
40	Dialog based Inspirational Instruction	I8
41	Practical Problem based Instruction	I9
42	Instrumental Procedure Oriented Instruction	I10
44	Non-directive Instruction	I12
45	Chemistry	Chem
46	Biology	Bio
47	Geo Subjects	Geo
48	Mathematics	Math
49	Physics	Phys
50	Complex Systems Technology	Cplx

5 ANOVA

5.1 Introduction ANOVA

In the first method, the analysis of variance (ANOVA) will be used. The research question chosen for this method is:

"Does students' work type interests significantly influence their level of interest in mathematics?"

The interest in mathematics subjects for the different students, will be used as the response variable. The interest in mathematics subjects is on a 'Likert scale' from 1 to 6, where 1 is 'No interest' and 6 is 'High interest'.

The work type interests will be used as the explanatory variable. The work type interests are variables with three levels: 2 = 'High liking', 1 = 'Medium liking', 0 = 'Low liking'. In order to answer the research question, only samples with high liking (2) will be used. The different work type variables are presented in table 2.

Table 2: Work type interests

No.	Work Type
W1	Practical People Oriented Service
W2	Practical Technical Work
W3	Creative Entrepreneur Work
W5	Idealistic Value Oriented Work
W6	Gross Motor Work
W8	Logical System Technical Work
W9	Communication and Informative Work
W11	Design Oriented Work
W12	Technical Computations
W14	Technical Problem Solving
W15	Development and Change Oriented Work
W16	Technical System Control
W17	Humanitarian Service Work
W19	Chaos and Change Oriented Work
W20	mathematicsematical Computations

5.2 Method section

ANOVA is used to analyze the difference between the means of several groups. The method is used to test the null hypothesis that the means of two or more populations are equal.

5.2.1 Model assumptions

In order to perform an ANOVA analysis, some assumptions has to be made. The model assumptions are as follows:

- Each observation should be independent of each other.
- There is a linear relationship between the response variable and the explanatory variable.
- The residuals are normally distributed.
- The residuals have constant variance (homoscedasticity).

For the research question chosen above, the model can be formulated as follows:

```
Yij = μ + αi + εij
```

where:

- Yij is the observed response for sample j in group i, here the response is the interest in mathematics subjects
- μ is the overall mean
- αi is the effect of the i-th work type
- εij is the random error term,ε ij ~ N(0, σ^2)

5.2.2 Hypotheses

The null and alternative hypotheses can be formulated as follows:

- Null hypothesis (H0): The means of the interest in mathematics are equal for all work type interests.

```
H0: μ1 = μ2 = μ3 = ... = μk
```

- Alternative hypothesis (H1): The means of the interest in mathematics are not equal for all work type interests.

```
H1: μi ≠ μj for at least one pair of groups
```

For the purpose of this assignment, the null hypothesis will be rejected if the p-value is less than 0.05.

The model will be fitted on the training data, and then used to predict the same response variable on the test data.

In the following section, the data will be wrangled and explored - before performing the ANOVA analysis.

5.2.3 Data wrangling

In order to perform the ANOVA analysis, the data has to be preprocessed and wrangled. The first data set used, is the `Train` data set, containing information about 5000 different students. The data set will be filtered to only include the variables of interest for the ANOVA analysis. The data set is converted to long format, where the work type interests are in one column and the interest in mathematics is in another column. The data set is filtered to only include samples with high liking (2) for the work type interests.

The test data will be wrangled in the same way.

Code chunks for data wrangling are presented below.

```
# select the variables of interest for the ANOVA analysis
math.data <- Train |>
  dplyr::select(Math, W1, W2, W3, W5, W6, W8,
              W9, W11, W12, W14, W15, W16, W17, W19, W20)
```

```
# select the same variables from the test data set
math.test <- Test |>
  dplyr::select(Math, W1, W2, W3, W5, W6, W8,
                W9, W11, W12, W14, W15, W16, W17, W19, W20)
```

```
# convert all work type variables into factors
math.data <- math.data |>
  mutate(across(W1:W20, as.factor))

# convert data to long format
math_long <- math.data %>%
  pivot_longer(
    cols = -Math,  # All columns except response
    names_to = "work_type",
    values_to = "interest_level"
  )

# Filter out samples with low liking (0) and medium liking (1)
math_high <- math_long %>%
  filter(interest_level == 2) %>%
  dplyr::select(-interest_level)
```

```
# convert all work type variables into factors
math.test <- math.test |>
  mutate(across(W1:W20, as.factor))

# convert data to long format
math_long_test <- math.test %>%
  pivot_longer(
    cols = -Math,  # All columns except response
    names_to = "work_type",
    values_to = "interest_level"
  )

# Filter out samples with low liking (0) and medium liking (1)
test.data <- math_long_test %>%
  filter(interest_level == 2) %>%
  dplyr::select(-interest_level)
```

5.2.4 Data visualization

```
# plot using ggplot
ggplot(math_high, aes(x = work_type, y = Math)) +
  geom_boxplot(aes(fill = work_type)) +
  facet_wrap(~ work_type, scales = "free_x") +
  labs(title = "Interest in mathematics by Work Type Interests",
       y = 'Interest in mathematics',
       x = 'Work type') +
  scale_fill_manual(values = rep("#F8BBD0", 15))  +
  theme_minimal() +
  theme(legend.position = 'none',
        title = element_text(size = 11))
```

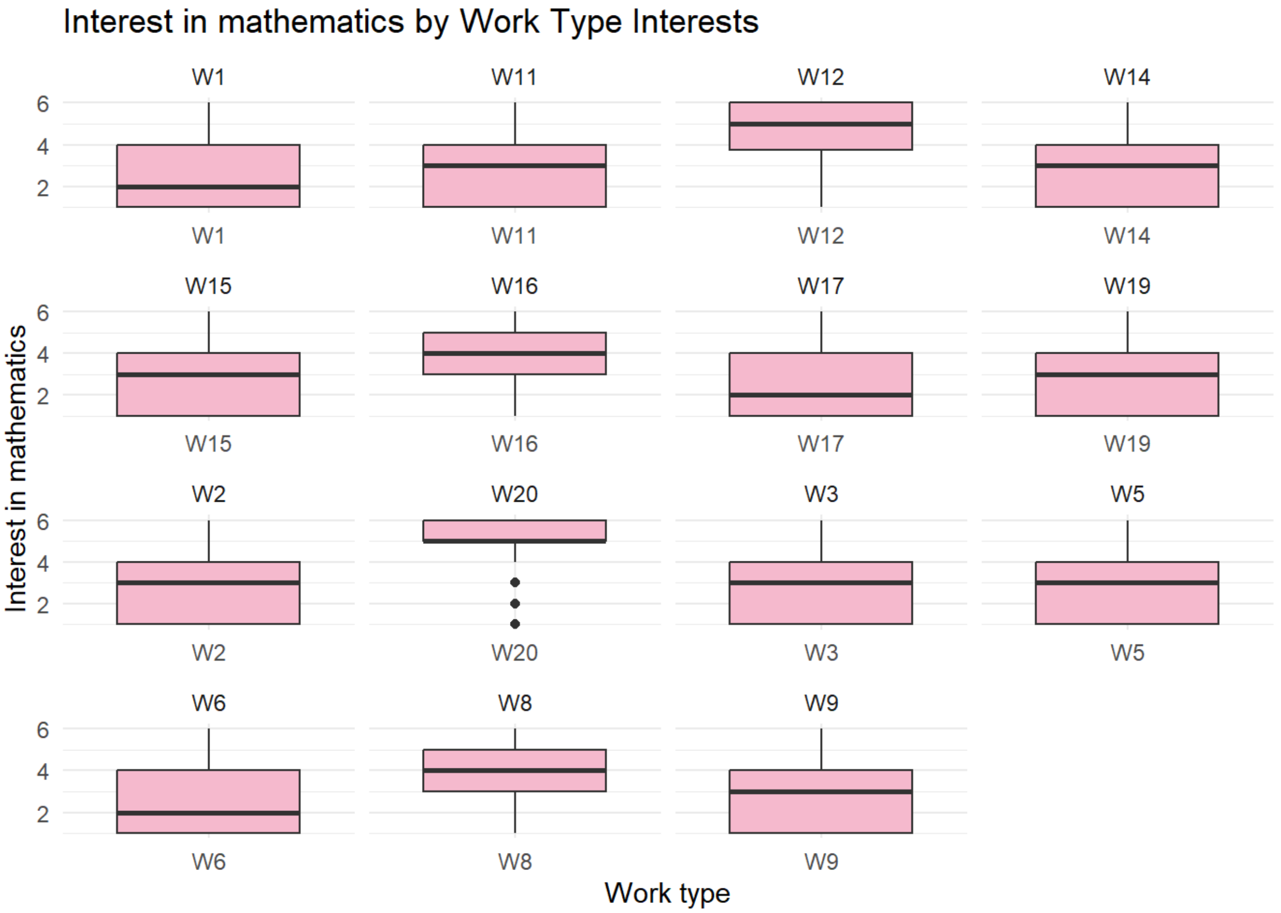

Figure 1: Boxplot of interest in mathematics by work type interests

Figure 1 shows the distribution of mathematics interest for different work type interests. The samples with a high interest in work type W12 (Technical Computations), W20 (mathematical computations) and W8 (Logical System Technical Work) have the highest interest in mathematics.

5.2.5 Modeling in R

The model was fitted using the `lm()` function in R, with `Math` as the response variable and `work_type` as the explanatory variable.

The data set used for analysis was `math_high`, which only contains samples with high liking (2) for the work type interests.

Contrasts was set to `contr.sum`, meaning that all work type variables are compared to the overall mean.

After fitting the linear model, model assumptions were checked using diagnostic plots. The ANOVA table was created using the `anova()` function. To predict the interest in mathematics for the test data, the `predict()` function was used.

The fitted model and its output is presented in the results section below.

5.3 Results

5.3.1 Fitting the model

```
# fit the model
mod1 <- lm(Math ~ work_type, data = math_high, contrasts = "contr.sum")
```

5.3.2 Check model assumptions

The model assumptions regarding normality and homoscedasticity for the residuals have to be checked. Model assumptions can be checked using diagnostic plots.

```
par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(mod1, which = c(1,2,4,5), add.smooth=TRUE)
```

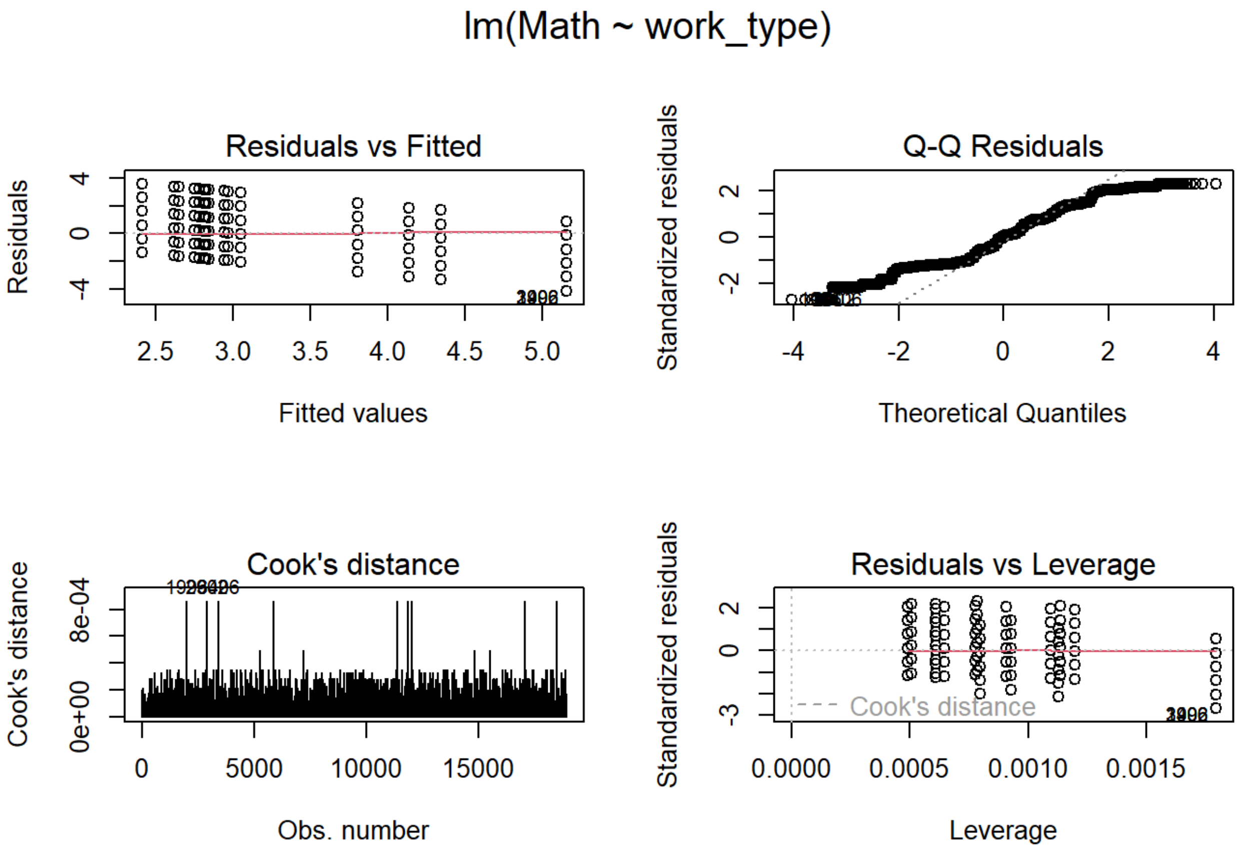

Figure 2: Diagnostic plots for the fitted model

5.3.2.1 Residuals vs Fitted

The residuals vs fitted plot shows the residuals on the y-axis and the fitted values on the x-axis. There is no strong curve pattern, which indicates that the linearity assumption is met. The residuals are fairly evenly spread across the fitted values, which indicates that the homoscedasticity assumption is met. There is a bit more spread on the lower end. The variance is not perfectly constant, but acceptable.

5.3.2.2 Q-Q Residuals

The Q-Q plot shows the quantiles of the residuals on the y-axis and the quantiles of a normal distribution on the x-axis. The points should follow a straight line if the residuals are normally distributed. The curve above has a slight S-shape, which indicates that the residuals are not perfectly normally distributed. There are deviations at both tails.

5.3.3 Performing ANOVA

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Math
##              Df Sum Sq Mean Sq F value    Pr(>F)
## work_type    14   7632  545.14  226.72 < 2.2e-16 ***
## Residuals 18901  45448    2.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the ANOVA table, work type interest has a significant effect on the interest in mathematics (p-value < 0.001). We reject the null hypothesis and conclude that there is a significant difference in the interest in mathematics for different work type interests.

Work_type has a sum of squares value equal to 7632, while the residuals have a sum of squares value equal to 45448. This means that there is a lot of unexplained variance in the model.

```
summod1 <- summary(mod1, cor=FALSE)
print(summod1)
```

```
##
## Call:
## lm(formula = Math ~ work_type, data = math_high, contrasts = "contr.sum")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1559 -1.6203  0.0556  1.1824  3.5865
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.65241    0.03833  69.207  < 2e-16 ***
## work_typeW11  0.10108    0.05773   1.751  0.08000 .
## work_typeW12  1.69331    0.06463  26.201  < 2e-16 ***
## work_typeW14  0.31917    0.06401   4.987 6.20e-07 ***
## work_typeW15  0.17533    0.05496   3.190  0.00142 **
## work_typeW16  1.15260    0.06084  18.945  < 2e-16 ***
## work_typeW17 -0.03209    0.05181  -0.619  0.53570
## work_typeW19  0.19241    0.06042   3.185  0.00145 **
## work_typeW2   0.13168    0.06482   2.032  0.04222 *
## work_typeW20  2.50350    0.07601  32.935  < 2e-16 ***
## work_typeW3   0.29203    0.05419   5.389 7.18e-08 ***
## work_typeW5   0.40141    0.06592   6.090 1.15e-09 ***
## work_typeW6  -0.23889    0.05796  -4.122 3.78e-05 ***
## work_typeW8   1.48544    0.05818  25.532  < 2e-16 ***
## work_typeW9   0.16523    0.05145   3.211  0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## s: 1.551 on 18901 degrees of freedom
## Multiple R-squared: 0.1438,
## Adjusted R-squared: 0.1431
## F-statistic: 226.7 on 14 and 18901 DF,  p-value: < 2.2e-16
```

The summary from the linear model is presented above. The table presents the coefficients for the different work type interests, and their p-values.

The mean mathematics interest for all work type interests is 2.65. The groups W12 (Technical Computations), W20 (mathematical computations) and W8 (Logical System Technical Work), all contribute significantly to the mathematics interest (positively). Which the boxplot also indicated.

W2 (Practical Technical Work), W11 (Design Oriented Work) and W17 (Humanitarian Service Work) are the non-significant work type interests, meaning that they do not contribute significantly to the interest in mathematical subjects.

The R-squared value is 0.1431, which means that 14.3% of the variance in the interest in mathematical subjects can be explained by the work type interests (this model)

5.3.4 Predicting mathematics interest on test data samples

In the code below, `mod1` (the fitted model) is used to predict the interest in mathematics for the samples in the test data.

```
predictions <- predict(mod1, newdata = test.data)
```

5.3.4.1 Calculate root mean squared error

To evaluate the models performance, the root mean squared error (RMSE) can be calculated. The formula and code is shown below.

```
RMSE <- sqrt(mean(predictions - test.data$Math)^2)
cat('The root mean squared value on the test data is:', RMSE)
```

```
## The root mean squared value on the test data is: 0.2231142
```

The model has a root mean squared error (RMSE) of 0.22. On average, the model misses the predicted value in mathematics interest by 0.22.

5.4 Conclusion

The model above was fitted using the training data, and used to predict the interest in mathematics for the test data. The diagnostic plots showed that the model assumptions were not perfectly met, especially the normality assumption.

In the ANOVA analysis, the work type interest was found to have a significant effect on the interest in mathematics.Looking at the summary from the linear model, the work type interests W12 (Technical Computations), W20 (mathematical computations) and W8 (Logical System Technical Work) all contribute significantly to the interest in mathematics. This makes sense, since these work type interests are somehow related to mathematics.

The work type interests W2 (Practical Technical Work), W11 (Design Oriented Work) and W17 (Humanitarian Service Work) are not significant, meaning that they do not contribute significantly to the interest in mathematical subjects. Their average is not significantly different from the overall mean. The work types W2 and W11 are more practical and creative, while W17 is more humanitarian, which *could* explain why they are not significant.

Using the model to predict the interest in mathematics for the test data, the root mean squared error (RMSE) was found to be 0.22. On average, the model misses the actual value in mathematics interest by 0.22.

Since the R-squared value in the linear model is only 0.1431, there is a lot of unexplained variance. For further analysis, it would be interesting to test other explanatory variables, such as the personality variables - and how they influence the interest in mathematical subjects.

6 Classification

6.1 Introduction classification

Classification is a supervised learning method used to predict the class membership of a sample. In this assignment, a classification method will be used to predict a students sex, based on their work type interests and instructional preferences.

The research question chosen for this method is:

"Is it possible to determine a student's sex based on their instructional preferences and work interests?"

The response variable is the categorical variable `Sex` , "male" or "female", converted into a binary variable (0 or 1).

The explanatory variables are the students' instructional preferences and work interests. The work type interests are presented in the ANOVA part of this assignment, and can be found in Table 2. The students' instructional preferences are:

Table 3: Instructional preferences

No.	Instructional preference
I1	Practical Inductive Instruction
I3	Abductive Theoretical Instruction
I4	Dialog and Group Oriented Theoretical Instruction
I6	Methodological Formula Oriented Instruction
I7	Explorative Inspirational Instruction
I8	Dialog based Inspirational Instruction
I9	Practical Problem based Instruction
I10	Instrumental Procedure Oriented Instruction
I12	Non-directive Instruction

Both work type interests and instructional preferences are categorical variables with three levels: 2 = 'High liking', 1 = 'Medium liking', 0 = 'Low liking'.

6.2 Method section

For classification, Linear Discriminant Analysis (LDA) will be used. LDA is a classification method used to find a linear combination of features that separates two or more classes. Given the research question above, the task is to separate male students from female students.

6.2.1 Model assumptions

The assumptions for LDA are as follows:

- The features are samples from a normal distribution.
- The features have equal variance-covariance structures across classes.

6.2.2 Data wrangling and preprocessing

The data set used for the classification analysis is the `Train` data set, as for the ANOVA analysis. In order to perform the classification analysis, the data has to be manipulated and preprocessed.

First of all, a data frame including the response variable and the explanatory variables has to be created for both the training and test data. Missing values have to be checked.

```
# extract the response and explanatory variables from the training data set
train.df <- Train |>
  dplyr::select(Sex, I1:I12, W1:W20)

# extract the same variables from the test data set
test.df <- Test |>
  dplyr::select(Sex, I1:I12, W1:W20)

# check for missing values
cat('There are', is.na(train.df) |> sum(),
    'missing values in the training data set and', is.na(test.df)
    |> sum(), 'missing values in the test data set.')
```

```
## There are 0 missing values in the training data set and 0 missing values in the test
data set.
```

The response variable must be converted into a factor, since it is a categorical variable. The `ifelse()` function in R is used to convert the response variable into a binary variable.

```
train.df$Sex <- ifelse(train.df$Sex == "M", 1, 0)
```

```
# count the occurrences of each class in the response variable
table(train.df$Sex)
```

```
##
##    0    1
## 3231 1769
```

The data is slightly unbalanced, with about 65% of the data in the female class (3231 samples).

6.2.3 Visualizing correlations

LDA assumes non-multicollinearity. The correlation between different variables gives an indication of multicollinearity, and how different variables are correlated.

```
# calculate the correlation matrix
ggcorr(train.df[-1]) +
  labs(title = "Correlation matrix for the explanatory variables") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_fill_gradient2(low = "plum2", mid = "white", high = "maroon3",
                       midpoint = 0, limit = c(-1, 1), name = "Correlation")
```

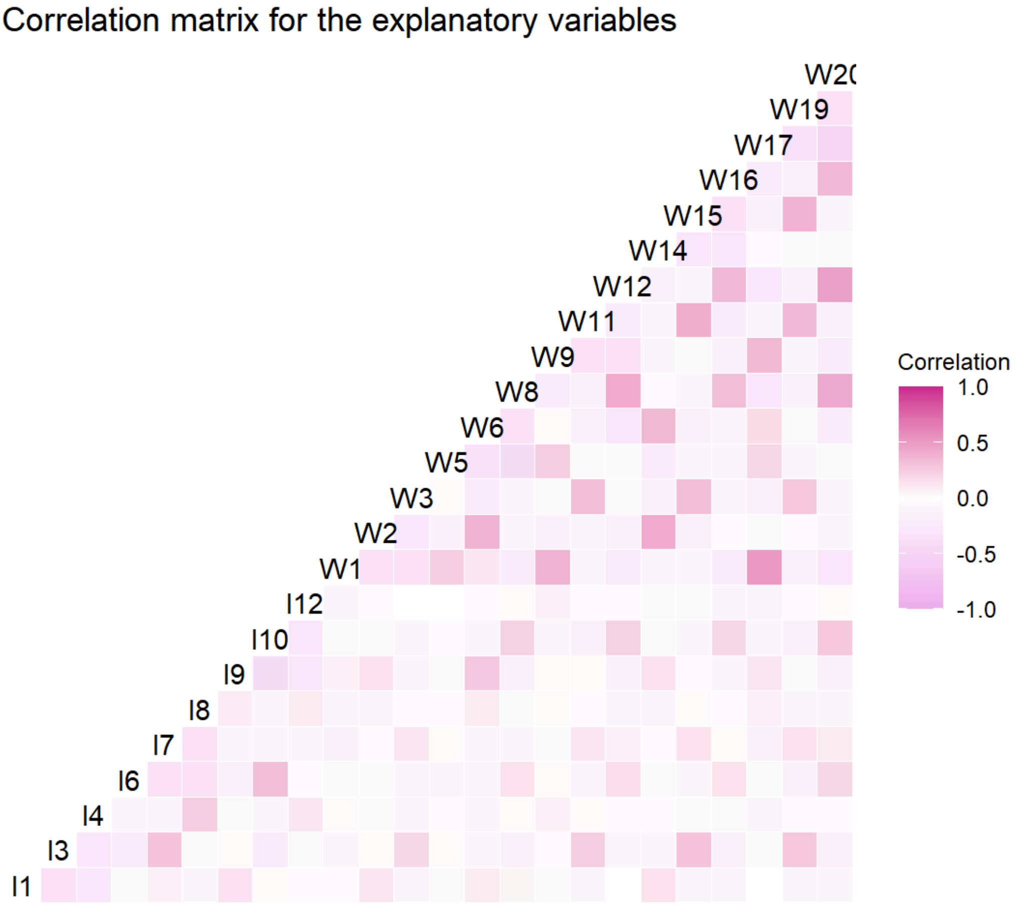

Figure 3: Correlation matrix for the explanatory variables

Non of the feature combinations looks highly correlated, but there are some moderate correlations. The highest correlation is between the features W1 and W17, around 0.5.

6.2.4 Modeling LDA in R

The LDA model is fitted using the `lda()` function in R from the `MASS` package. The response variable is `Sex` , and the explanatory variables are the work type interests and instructional preferences. The model is fitted on the training data set, `train.df` .

For prior probabilities, the proportions of the two classes in the training data set are used. Since the female class accounts for 65% of the samples in the data set, the prior is set to 0.65 for class 0 (female) and 0.35 for class 1 (male).

The model was used to predict the Sex for the test data set, `test.df` . A confusion matrix was created to evaluate the model performance on the test data.

6.3 Results

6.3.1 Fitting the model

```
mod2 <- lda(Sex ~ W1 + W2 + W3 + W5 + W6 + W8 + W9 + W11 + W12 + W14 + W15 +
            W16 + W17 + W19 + W20 + I1 + I3 + I4 + I6 + I7 + I8 + I9 + I10 +
            I12, data = train.df, prior = c(0.65, 0.35))
mod2
```

```
## Call:
## lda(Sex ~ W1 + W2 + W3 + W5 + W6 + W8 + W9 + W11 + W12 + W14 +
##     W15 + W16 + W17 + W19 + W20 + I1 + I3 + I4 + I6 + I7 + I8 +
##     I9 + I10 + I12, data = train.df, prior = c(0.65, 0.35))
##
## Prior probabilities of groups:
##    0    1
## 0.65 0.35
##
## Group means:
##         W1       W2       W3        W5       W6        W8        W9      W11
## 0 1.280409 0.815537 1.239864 0.9399567 1.002167 0.8585577 1.3574745 1.142061
## 1 0.669870 1.053703 1.192764 0.6167326 1.036179 1.1413228 0.9994347 1.075184
##         W12       W14      W15       W16       W17      W19       W20        I1
## 0 0.8424636 0.7790158 1.130300 0.8625812 1.3967812 1.016404 0.4258743 0.8622717
## 1 0.9853024 1.0480497 1.166761 1.0486150 0.9315998 1.133974 0.6992651 0.9542114
##          I3       I4       I6        I7        I8        I9       I10       I12
## 0 0.7465181 1.122253 1.272052 0.7638502 0.7879913 0.9665738 0.8901269 1.0417827
## 1 0.8479367 1.074053 1.169022 0.9694743 0.7925382 0.9304692 0.9417750 0.9621255
##
## Coefficients of linear discriminants:
##             LD1
## W1  -0.96397910
## W2  -0.37386297
## W3  -0.65279490
## W5  -0.19721033
## W6  -0.01864300
## W8   0.02816457
## W9  -0.41660213
## W11 -0.57702046
## W12 -0.34570519
## W14  0.56323174
## W15  0.29924299
## W16  0.32148694
## W17 -0.21879230
## W19  0.23080155
## W20  0.14486265
## I1   0.20184324
## I3   0.29634174
## I4   0.08902618
## I6  -0.10144098
## I7   0.11688669
## I8   0.11614072
## I9  -0.17054551
## I10 -0.08013014
## I12 -0.22529145
```

The output from the fitted model, shows the group means for the different explanatory variables. This is the difference in means between female and male students for the different work type interests and instructional preferences. For example W17 (Humanitarian Service Work) has a group mean of 1.40 for female, and 0.93 for males. Big differences in means indicate that the variable is important for classification.

The coefficients for the first linear discriminant component (LD1) is also presented in the output. This is the weights for the different explanatory variables. Big weights indicate that the variable is important for classification, and small weights indicate that the variable is not important.

Variables with small weights were removed from the model to improve the model performance and reduce model complexity. The reduced model is shown in the code chunk below.

```
mod.red <- lda(Sex ~ W1 + W2 + W3 + W5 + W9 + W11 + W12 + W14 + W15 +
            W16 + W17 + W19 + W20 + I1 + I3 + I6 + I7 + I8 + I9 +
            I10 + I12, data = train.df, prior = c(0.65, 0.35))
```

6.3.2 Predictions on the test data

```
Predictions_test <- predict(mod.red, newdata = test.df)
```

6.3.3 Evaluate the model performance using a confusion matrix

```
conf.matrix_test <- confusion(test.df$Sex, Predictions_test$class)
```

```
##          True
## Predicted    F    M
##   0       3104  704
##   1        420  772
##   Total   3524 1476
##   Correct 3104  772
##
## Proportions correct
##         F         M
## 0.8808173 0.5230352
##
## N correct/N total = 3876/5000 = 0.7752
```

The confusion matrix shows an accuracy of 0.7752, meaning that the model correctly classifies 78% of the samples in the test data set. The proportions of correct predictions are higher in the female class (88%) than in the male class (52%).

The apparent error rate (APER) can also be considered. For female, this rate is 420/3524 = 0.12. For male, the APER is 704/1476 = 0.48.

To further evaluate model performance, the F1-score can be calculated. F1-score is the harmonic mean of precision and recall, and is a good measure of model performance when the data is unbalanced.

F1-score for the male class:

```
# Extract values from confusion matrix
TP_male <- 772  # True positives (Predicted 1 and Actual 1)
TN_male <- 3104  # True negatives (Predicted 0 and Actual 0)
FP_male <- 704  # False positives (Predicted 0 and Actual 1)
FN_male <- 420  # False negatives (Predicted 1 and Actual 0)

# Calculate Precision and Recall
Precision_male <- TP_male / (TP_male + FP_male)
Recall_male <- TP_male / (TP_male + FN_male)

# Calculate F1 score
F1_score_male <-
  2 * (Precision_male * Recall_male) /
  (Precision_male + Recall_male)

# Print the F1 score
cat('The F1_score for the male class is:', F1_score_male)
```

```
## The F1_score for the male class is: 0.5787106
```

F1-score for the female class:

```
# Extract values from confusion matrix
TP_female <- 3104  # True positives (Predicted 0 and Actual 0)
TN_female <- 772  # True negatives (Predicted 1 and Actual 1)
FP_female <- 420  # False positives (Predicted 1 and Actual 0)
FN_female <- 704  # False negatives (Predicted 0 and Actual 1)

# Calculate Precision and Recall
Precision_female <- TP_female / (TP_female + FP_female)
Recall_female <- TP_female / (TP_female + FN_female)

# Calculate F1 score
F1_score_female <-
  2 * (Precision_female * Recall_female) /
  (Precision_female + Recall_female)

# Print the F1 score
cat('The F1_score for the female class is:', F1_score_female)
```

```
## The F1_score for the female class is: 0.8466994
```

The macro average F1-score is calculated as the average of the F1-scores for the two classes:

```
macro_F1 <- (F1_score_female + F1_score_male) /2
cat('The macro average F1-score is:', macro_F1)
```

```
## The macro average F1-score is: 0.712705
```

6.4 Conclusion

The classification method used in this assignment was Linear Discriminant Analysis (LDA). The average accuracy of the model was found to be 0.7752, meaning that the model correctly classifies 78% of the samples in the test data set. The proportion of correct predictions was best for the female class, maybe due to its high abundance.

Since the classes are slightly unbalanced, the F1-score can give a better indication of model performance. The macro average F1-score was 0.71. Treating the classes equally, the model has decent performance.

For further analysis, it would be interesting to test other classification methods, such as Random Forest or Support Vector Machines (SVM).

To conclude, work type interest and instruction preferences can - to some extent - be used to predict the Sex of students, but the model is not perfect.

7 Principal Component Analysis

7.1 Introduction PCA

Principal Component Analysis (PCA) is an unsupervised learning method used to reduce the dimensionality of a data set. In this assignment, PCA will be used to reduce the dimensionality of the data set, specifically students instructional preferences. The research question chosen for this method is:

"How do students' instructional preferences relate to each other?"

The variables used for PCA are the students' instructional preferences (I1-I12), presented in Table 3.

7.2 Method section

PCA is a method used to reduce the dimensionality of a data set by transforming the original variables into a new set of uncorrelated variables, called principal components.

The principal components are linear combinations of the original variables, and they are ordered by the amount of variance they explain in the data set.

The first principal component explains the most variance, the second principal component explains the second most variance, and so on.

7.2.1 Model assumptions

The assumptions for PCA are as follows:

- The relationship between the variables is linear.
- Components with high variance are more important than components with low variance.
- Components are orthogonal to each other.
- The data is standardized, meaning that all variables have the same scale. If the data is not standardized, PCA will give to much weight to variables with large scales.

7.2.2 Data wrangling

The data used for PCA is the `Train` data set.

The data set is filtered to only include the instructional preferences (I1-I12), saved in the data frame `instruction_prefs`.

```
instruction_prefs <- Train[, c("I1", "I3", "I4", "I6", "I7",
                               "I8", "I9", "I10", "I12")]
```

7.2.3 Visualize data distribution

```
# transform the data to long format
instruction_prefs_long <-
  instruction_prefs |>
  pivot_longer(
    cols = everything(),  # All columns
    names_to = "instructional_preference",
    values_to = "liking") |>
  group_by(instructional_preference, liking) |>
  summarise(count = n())

instruction_prefs_long$liking <- factor(instruction_prefs_long$liking)

# plot using ggplot
ggplot() +
  geom_col(data = instruction_prefs_long,
           aes(x = instructional_preference, y = count, fill = liking)) +
  labs(title = "Distribution of Instructional Preferences",
       y = 'Count',
       x = 'Instructional preference') +
  theme_minimal() +
  theme(title = element_text(size = 11, hjust = 0.5)) +
  scale_fill_manual(values = c("0" = "#FFB6C1",
                               "1" = "plum",
                               "2" = "maroon"))
```

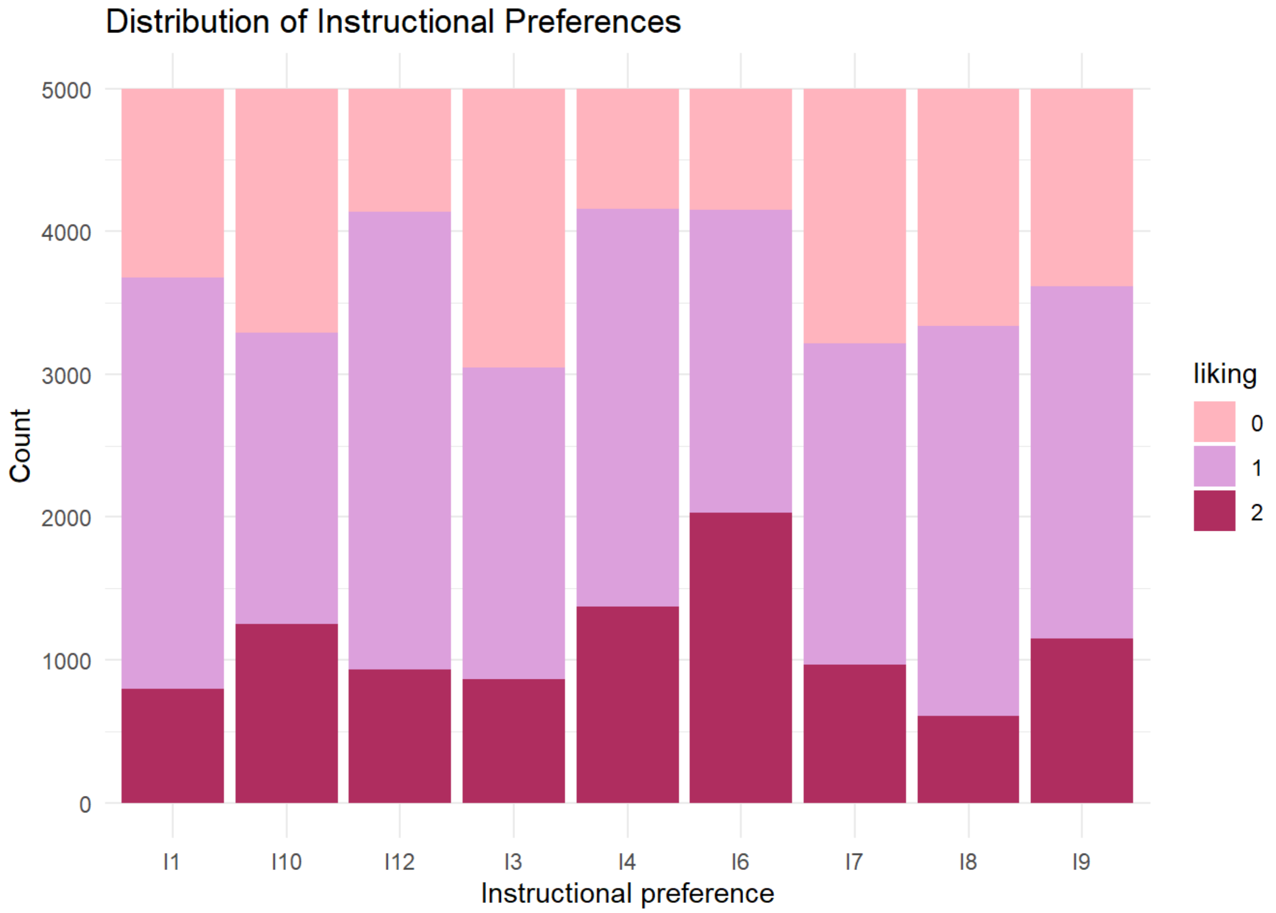

Figure 4: Bar plot of the distribution of instructional preferences

The bar plot shows the distribution of the different instructional preferences. The distribution between the different preferences is quite even, but there are some preferences that are more popular than others. For example, the dark red color is quite abundant in I6 (Methodological Formula Oriented Instruction), indicating that many students have a high liking for this instruction type.

7.3 Modeling

The PCA model is fitted using the `prcomp()` function in R. The data is standardized using the `scale.` argument.

The model is fitted on the training data set, `instruction_prefs` , and can be seen in the results section below.

7.4 Results

```
pca <- prcomp(instruction_prefs, scale. = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.3663 1.3141 1.1962 1.0787 1.0054 0.9718 0.58186
## Proportion of Variance 0.2074 0.1919 0.1590 0.1293 0.1123 0.1049 0.03762
## Cumulative Proportion  0.2074 0.3993 0.5583 0.6876 0.7999 0.9048 0.94246
##                            PC8     PC9
## Standard deviation     0.51594 0.50171
## Proportion of Variance 0.02958 0.02797
## Cumulative Proportion  0.97203 1.00000
```

The PCA model above shows the proportion of variance explained by each principal component. The data is split into 9 principal components. The first principal component explains 21% of the variance, the second principal component explains 19% of the variance, and so on. To explain 90% of the variance, 6 principal components are needed.

The two first principal components can be visualized in a biplot, which shows the relationship between the original variables and the principal components.

7.4.1 Loadings for the first two principal components

```
loadings <- pca$rotation[, 1:2]
loadings
```

```
##             PC1          PC2
## I1  -0.07171609 -0.116811616
## I3   0.34911627 -0.375520165
## I4   0.10471066  0.483398779
## I6  -0.57360265 -0.006132085
## I7   0.28720982 -0.512433778
## I8   0.21605303  0.531403949
## I9   0.29433262  0.099314022
## I10 -0.55043983 -0.081837313
## I12  0.11918987  0.223804637
```

The loadings plot shows the relationship between the original variables and the first two principal components. The loadings are the coefficients for the linear combinations of the original variables. Loadings that are similar in sign and magnitude indicate that the variables are correlated. Loadings close to 1 or -1 indicate that the variable is important for the principal component.

In PC1, I6 (Methodological Formula Oriented Instruction) and I10 (Instrumental Procedure Oriented Instruction) are the most important variables, with loadings close to -1. These may be similar to each other. I7 (Explorative Inspirational Instruction) and I3 (Abductive Theoretical Instruction) also have similar loadings, and may share similar information.

In PC2, I4 (Dialog and Group Oriented Theoretical Instruction) and I8 (Dialog based Inspirational Instruction) are the most important variables, with loadings close to 1. These may also be similar to each other, based on similar loadings and direction.

7.4.2 Visualize PC1 and PC2

```
loadingplot(pca,  comps = c(1,2),
            scatter = TRUE, labels = 'names')
```

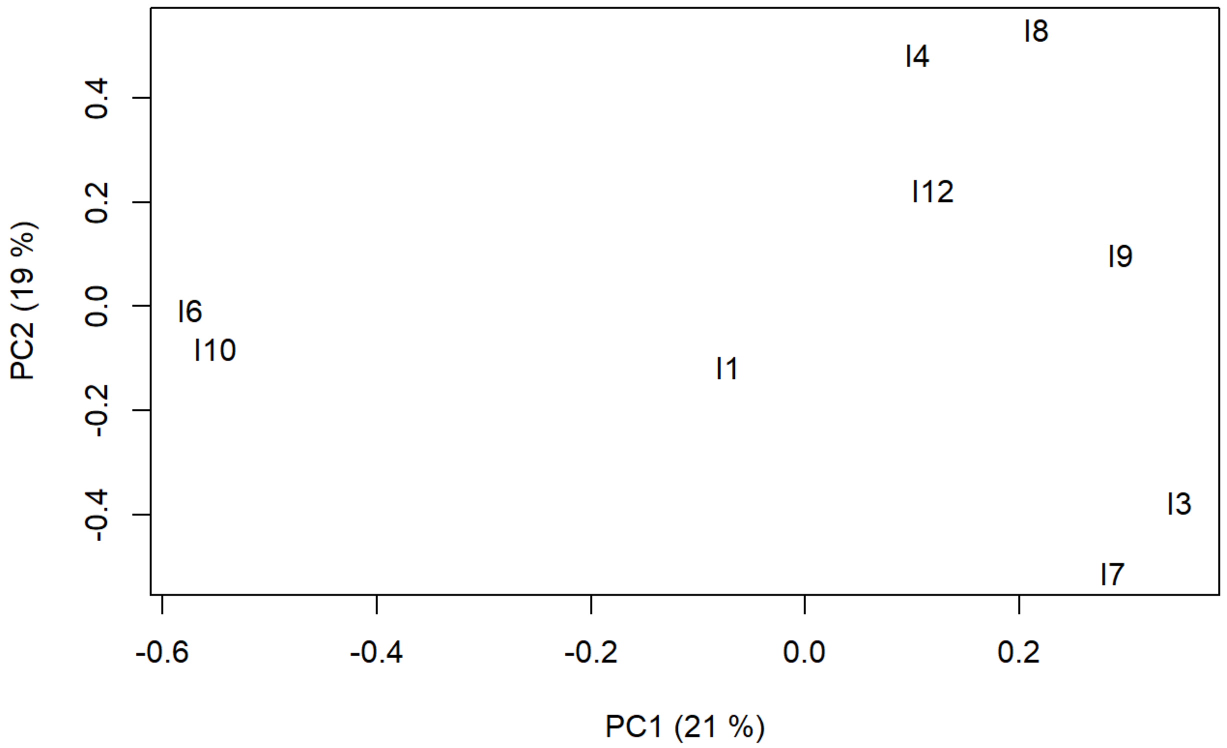

Figure 5: Pairplot of the first two principal components

The loading plot for the PCA, shows the relationship between the original variables and the first two principal components. There are some distinct clusters in the plot, indicating that some variables are similar to each other, especially for I6 and I10 - mentioned above. I7 and I3, and I4 and I8, are also close to each other indicating that they are similar.

7.5 Scores

```
pca_scores <- as.data.frame(pca$x[, 1:4])
head(pca_scores)
```

```
##          PC1          PC2        PC3        PC4
## 1 -1.8498097 -0.125140793 -0.7985922 -1.2443849
## 2 -0.5715165  1.465428760  0.3530464  1.3606480
## 3  1.1424186  0.498745293 -0.2288295  0.6695359
## 4 -0.3474928  0.007260797 -0.1664815  1.9977154
## 5  1.2881190 -0.261901606  1.1355167  0.5361014
## 6 -1.0033677  0.397775045  0.4054218 -0.2029233
```

The PCA scores show the coordinates of the samples in the new PCA space. Samples with similar coordinates are similar to each other. In this case, students with similar instructional preferences are close to each other in the PCA space.

```
# Scatter plot of the first two PCs
ggplot(pca_scores, aes(x = PC1, y = PC2)) +
  geom_point() +
  labs(title = "Scores plot of the first two PCs",
       x = "PC1", y = "PC2")
```

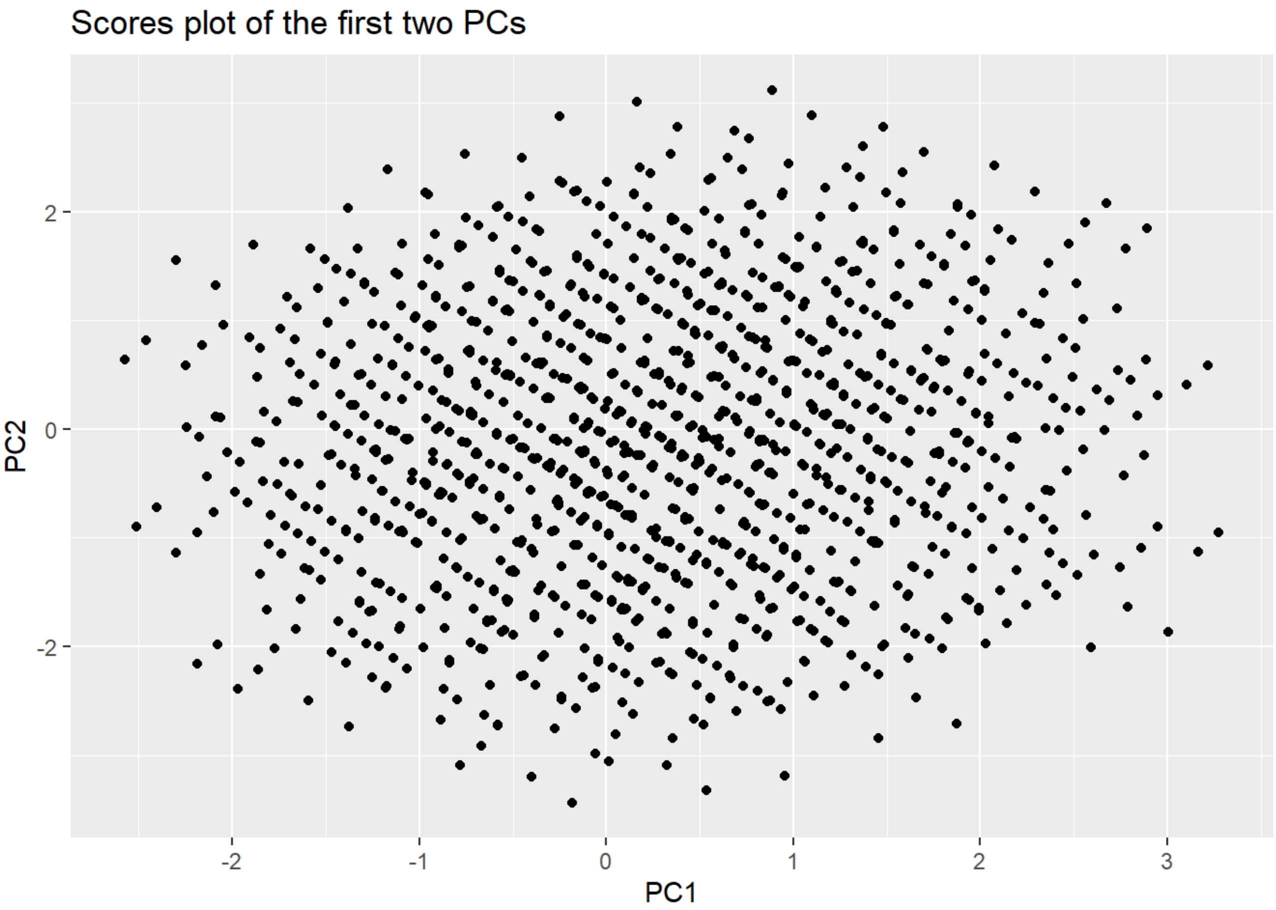

Figure 6: Scores plot of the first two principal components

Plotting the PCA scores shows the distribution of the samples in the PCA space. This plot doesn’t show any distinct clusters, it all looks quite random.

```
score.tbl <- as.data.frame(pca$x) |>
  bind_cols(instruction_prefs)

# find the relative variance explained by each component
relative_evar <- round(pca$sdev^2 / sum(pca$sdev^2), 2)

# plot a score plot
ggplot() +
  geom_text(data = score.tbl, mapping = aes(x = PC1, y = PC2, label = I10)) +
  labs(x = str_c('PC1 (', relative_evar[1], ')'),
       y = str_c('PC2 (', relative_evar[2], ')'),
       title = 'Score plot labeled by interest in I10')
```

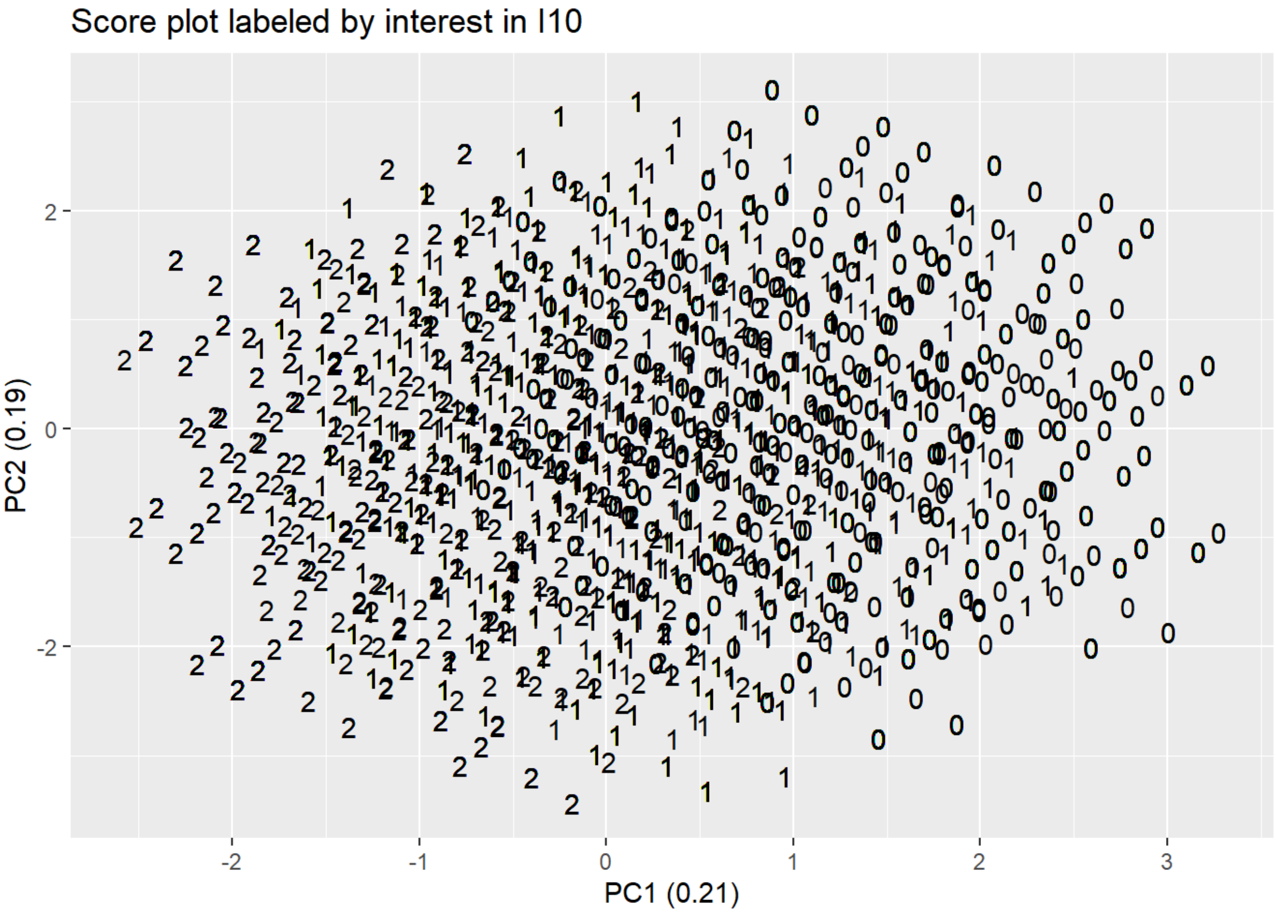

Figure 7: Score plot labeled by interest in I10

Figure 7 shows the score plot labeled by students interest in I10 (Instrumental Procedure Oriented Instruction). We can see that students with a high interest in I10 are clustered together in the left part, while students with no interest in I10 are clustered together in the right part. The students with moderate interest are placed in the middle.

```
score.tbl <- as.data.frame(pca$x) |>
  bind_cols(instruction_prefs)

# find the relative variance explained by each component
relative_evar <- round(pca$sdev^2 / sum(pca$sdev^2), 2)

# plot a score plot
ggplot() +
  geom_text(data = score.tbl, mapping = aes(x = PC1, y = PC2, label = I6)) +
  labs(x = str_c('PC1 (', relative_evar[1], ')'),
       y = str_c('PC2 (', relative_evar[2], ')'),
       title = 'Score plot labeled by interest in I6')
```

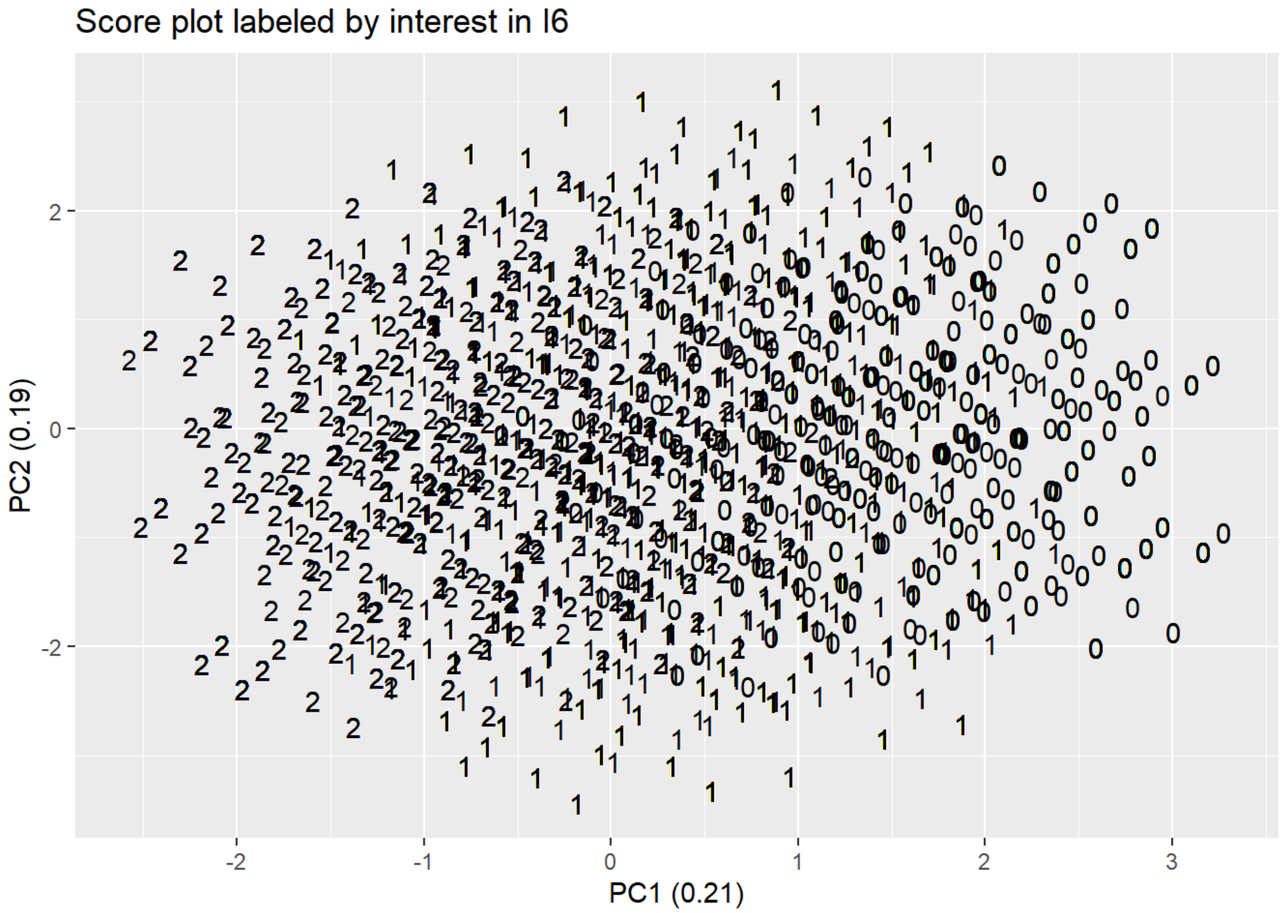

Figure 8: Score plot labeled by interest in I6

Figure 8 shows the score plot labeled by students interest in I6 (Methodological Formula Oriented Instruction). The same pattern as for figure 7 is seen here, students with a high interest in I6 are clustered together in the left part of the plot. Students with no interest is placed in the right part of the plot.

Similar score plots for the two instructional preferences I6 and I10, indicate that they are similar to each other (which we also saw in the loading plot) - the students interested/not interested in I6 often have the same preference for I10.

To further investigate the relationship between the different instructional preferences, a correlation matrix can be created.

```
# calculate the correlation matrix
cor_matrix <- cor(instruction_prefs)
cor_matrix
```

```
##              I1          I3          I4          I6          I7          I8
## I1   1.00000000 -0.34317080 -0.30930238 -0.06393524  0.06029707 -0.08055339
## I3  -0.34317080  1.00000000 -0.27707274 -0.22125000  0.30043109 -0.05989823
## I4  -0.30930238 -0.27707274  1.00000000 -0.11012825 -0.08082538  0.24984702
## I6  -0.06393524 -0.22125000 -0.11012825  1.00000000 -0.36983328 -0.34259899
## I7   0.06029707  0.30043109 -0.08082538 -0.36983328  1.00000000 -0.37777441
## I8  -0.08055339 -0.05989823  0.24984702 -0.34259899 -0.37777441  1.00000000
## I9   0.13358437 -0.03260720  0.02155683 -0.16343634 -0.08789023  0.08327736
## I10 -0.03228588 -0.22058746 -0.09239703  0.32969904 -0.10965651 -0.09461288
## I12 -0.03061601 -0.04238399  0.11892719 -0.02688447 -0.06832833  0.09773516
##              I9         I10         I12
## I1   0.13358437 -0.03228588 -0.03061601
## I3  -0.03260720 -0.22058746 -0.04238399
## I4   0.02155683 -0.09239703  0.11892719
## I6  -0.16343634  0.32969904 -0.02688447
## I7  -0.08789023 -0.10965651 -0.06832833
## I8   0.08327736 -0.09461288  0.09773516
## I9   1.00000000 -0.39606893 -0.25728663
## I10 -0.39606893  1.00000000 -0.29482089
## I12 -0.25728663 -0.29482089  1.00000000
```

The correlation matrix shows that I6 and I10 are moderately correlated (0.33). I3 and I7 are also moderately correlated (0.30), and I4 and I8 with a value of 0.25.

7.6 Conclusion

The PCA model was fitted using the training data set, and the first two principal components were visualized. The PCA model explained 90% of the variance in the data set using 6 principal components.

The two first principal components explained 21% and 19% of the variance, respectively - in total 40%.

The loadings showed that some instructional preferences may share similar information, especially:

- I6 (Methodological Formula Oriented Instruction) and I10 (Instrumental Procedure Oriented Instruction)
- I3 (Abductive Theoretical Instruction) and I7 (Explorative Inspirational Instruction)
- I4 (Dialog and Group Oriented Theoretical Instruction) and I8 (Dialog based Inspirational Instruction)

The scorings plot showed that students with a high interest in I6 and I10 are clustered together in the PCA space, which further indicates that these two instructional preferences are similar to each other.

The correlation matrix also showed correlation between these three pairs.

Looking into the instructional preferences, we can see that all the correlated pairs are somehow similar to each other, and the PCA model was able to capture this. It makes sense that students with high liking to one instructional preference, also likes similar instructional preferences.

PCA can give insights into the relationship between different variables, and can be used to reduce the dimensionality of a data set. For further analysis, it would be interesting to perform clustering on the PCA scores, to see if there are any distinct clusters in the reduced data set.

8 References

1. Sæbø S, Almøy T, Brovold H. Does academia disfavor contextual and extraverted students? Uniped. 11. november 2015;38(4):274–83.