Project Academic Performance Analysis

Deni Permana - Diaz Fahreza Akbar

2024-11-24

0. Import Library

Library yang akan digunakan

```
library(readr)  # Untuk membaca data dengan cepat dari berbagai format seperti CSV
library(tidyverse)  # Sekumpulan paket terintegrasi untuk analisis data yang mendukung tidy data
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)  # Untuk manipulasi data dengan sintaks yang bersih dan intuitif
library(caret)  # Untuk membangun model prediktif, termasuk pemilihan fitur dan evaluasi model
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(randomForest)  # Untuk membangun model Random Forest untuk klasifikasi dan regresi
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(rpart)  # Untuk membuat model pohon keputusan (decision trees)
library(cluster)  # Menyediakan fungsi untuk analisis clustering seperti k-means dan agglomerative
library(factoextra)  # Memudahkan visualisasi hasil analisis multivarian seperti clustering dan PCA
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(corrplot)  # Memfasilitasi visualisasi matriks korelasi
```

```
## corrplot 0.95 loaded
```

```
library(ggplot2)  # Paket grafis yang powerful untuk membuat visualisasi data
library(reshape2)  # Memudahkan mengubah data antara format lebar dan panjang
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(stats)  # Paket bawaan R untuk analisis statistik dasar
library(NbClust)  # Memfasilitasi penentuan jumlah cluster optimal dengan berbagai metode
library(viridis)  # Menawarkan palet warna yang baik untuk visualisasi yang dapat diakses
```

```
## Loading required package: viridisLite
```

1. Import Dataset

Dataset yang akan digunakan

```
# 1. Student Mat
student_mat_data <- read_delim("resources/dataset/student-mat.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 395 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# 2. Student Por
student_por_data <- read_delim("resources/dataset/student-por.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 649 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2. Exploration Data

```r
# 1. Display Data
head(student_mat_data, 5)
```

```
## # A tibble: 5 x 33
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob    Fjob     reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>   <chr>    <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home teacher  course
## 2 GP     F        17 U       GT3     T           1     1 at_home other    course
## 3 GP     F        15 U       LE3     T           1     1 at_home other    other
## 4 GP     F        15 U       GT3     T           4     2 health  services home
## 5 GP     F        16 U       GT3     T           3     3 other   other    home
## # i 22 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
## #   failures <dbl>, schoolsup <chr>, famsup <chr>, paid <chr>,
## #   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
## #   romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>, Dalc <dbl>,
## #   Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

```r
head(student_por_data, 5)
```

```
## # A tibble: 5 x 33
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob    Fjob     reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>   <chr>    <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home teacher  course
## 2 GP     F        17 U       GT3     T           1     1 at_home other    course
## 3 GP     F        15 U       LE3     T           1     1 at_home other    other
## 4 GP     F        15 U       GT3     T           4     2 health  services home
## 5 GP     F        16 U       GT3     T           3     3 other   other    home
## # i 22 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
## #   failures <dbl>, schoolsup <chr>, famsup <chr>, paid <chr>,
## #   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
## #   romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>, Dalc <dbl>,
## #   Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

```r
# 2. Display Data Structure
str(student_mat_data)
```

```
## spc_tbl_ [395 x 33] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ school    : chr [1:395] "GP" "GP" "GP" "GP" ...
##  $ sex       : chr [1:395] "F" "F" "F" "F" ...
##  $ age       : num [1:395] 18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr [1:395] "U" "U" "U" "U" ...
##  $ famsize   : chr [1:395] "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr [1:395] "A" "T" "T" "T" ...
##  $ Medu      : num [1:395] 4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : num [1:395] 4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr [1:395] "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr [1:395] "teacher" "other" "other" "services" ...
##  $ reason    : chr [1:395] "course" "course" "other" "home" ...
##  $ guardian  : chr [1:395] "mother" "father" "mother" "mother" ...
##  $ traveltime: num [1:395] 2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : num [1:395] 2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : num [1:395] 0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr [1:395] "yes" "no" "yes" "no" ...
##  $ famsup    : chr [1:395] "no" "yes" "no" "yes" ...
##  $ paid      : chr [1:395] "no" "no" "yes" "yes" ...
##  $ activities: chr [1:395] "no" "no" "no" "yes" ...
##  $ nursery   : chr [1:395] "yes" "no" "yes" "yes" ...
##  $ higher    : chr [1:395] "yes" "yes" "yes" "yes" ...
##  $ internet  : chr [1:395] "no" "yes" "yes" "yes" ...
##  $ romantic  : chr [1:395] "no" "no" "no" "yes" ...
##  $ famrel    : num [1:395] 4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : num [1:395] 3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : num [1:395] 4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : num [1:395] 1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : num [1:395] 1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : num [1:395] 3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : num [1:395] 6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : num [1:395] 5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : num [1:395] 6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : num [1:395] 6 6 10 15 10 15 11 6 19 15 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   age = col_double(),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
##   ..   studytime = col_double(),
```

```
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   absences = col_double(),
##   ..   G1 = col_double(),
##   ..   G2 = col_double(),
##   ..   G3 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(student_por_data)
```

```
## spc_tbl_ [649 x 33] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ school    : chr [1:649] "GP" "GP" "GP" "GP" ...
##  $ sex       : chr [1:649] "F" "F" "F" "F" ...
##  $ age       : num [1:649] 18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr [1:649] "U" "U" "U" "U" ...
##  $ famsize   : chr [1:649] "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr [1:649] "A" "T" "T" "T" ...
##  $ Medu      : num [1:649] 4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : num [1:649] 4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr [1:649] "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr [1:649] "teacher" "other" "other" "services" ...
##  $ reason    : chr [1:649] "course" "course" "other" "home" ...
##  $ guardian  : chr [1:649] "mother" "father" "mother" "mother" ...
##  $ traveltime: num [1:649] 2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : num [1:649] 2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : num [1:649] 0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr [1:649] "yes" "no" "yes" "no" ...
##  $ famsup    : chr [1:649] "no" "yes" "no" "yes" ...
##  $ paid      : chr [1:649] "no" "no" "no" "no" ...
##  $ activities: chr [1:649] "no" "no" "no" "yes" ...
##  $ nursery   : chr [1:649] "yes" "no" "yes" "yes" ...
##  $ higher    : chr [1:649] "yes" "yes" "yes" "yes" ...
##  $ internet  : chr [1:649] "no" "yes" "yes" "yes" ...
##  $ romantic  : chr [1:649] "no" "no" "no" "yes" ...
##  $ famrel    : num [1:649] 4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : num [1:649] 3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : num [1:649] 4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : num [1:649] 1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : num [1:649] 1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : num [1:649] 3 3 3 5 5 5 3 1 1 5 ...
```

```
##  $ absences  : num [1:649] 4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : num [1:649] 0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : num [1:649] 11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : num [1:649] 11 11 12 14 13 13 13 13 17 13 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   age = col_double(),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
##   ..   studytime = col_double(),
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   absences = col_double(),
##   ..   G1 = col_double(),
##   ..   G2 = col_double(),
##   ..   G3 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# 3. Descriptive Stats
summary(student_mat_data)
```

```
##     school              sex                 age         address
##  Length:395         Length:395         Min.   :15.0   Length:395
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :17.0   Mode  :character
##                                        Mean   :16.7
##                                        3rd Qu.:18.0
##                                        Max.   :22.0
##    famsize            Pstatus               Medu            Fedu
##  Length:395         Length:395         Min.   :0.000   Min.   :0.000
```

```
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :3.000   Median :2.000
##                                        Mean   :2.749   Mean   :2.522
##                                        3rd Qu.:4.000   3rd Qu.:3.000
##                                        Max.   :4.000   Max.   :4.000
##      Mjob               Fjob              reason            guardian
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    traveltime      studytime        failures       schoolsup
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :2.000   Median :0.0000   Mode  :character
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##     famsup              paid           activities          nursery
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     higher            internet           romantic             famrel
##  Length:395         Length:395         Length:395         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.944
##                                                           3rd Qu.:5.000
##                                                           Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health         absences            G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

```r
summary(student_por_data)
```

```
##     school              sex                 age          address
##  Length:649         Length:649         Min.   :15.00   Length:649
##  Class :character   Class :character   1st Qu.:16.00   Class :character
##  Mode  :character   Mode  :character   Median :17.00   Mode  :character
##                                        Mean   :16.74
##                                        3rd Qu.:18.00
##                                        Max.   :22.00
##    famsize            Pstatus              Medu            Fedu
##  Length:649         Length:649         Min.   :0.000   Min.   :0.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :2.000   Median :2.000
##                                        Mean   :2.515   Mean   :2.307
##                                        3rd Qu.:4.000   3rd Qu.:3.000
##                                        Max.   :4.000   Max.   :4.000
##      Mjob               Fjob               reason            guardian
##  Length:649         Length:649         Length:649         Length:649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    traveltime      studytime        failures       schoolsup
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:649
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :2.000   Median :0.0000   Mode  :character
##  Mean   :1.569   Mean   :1.931   Mean   :0.2219
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##     famsup              paid            activities          nursery
##  Length:649         Length:649         Length:649         Length:649
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    higher            internet           romantic            famrel
##  Length:649         Length:649         Length:649         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.931
##                                                           3rd Qu.:5.000
##                                                           Max.   :5.000
##     freetime        goout            Dalc            Walc          health
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000
##  Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median :4.000
##  Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536
##  3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000
##  Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000
##     absences            G1              G2              G3
##  Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
```

```
##  1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
##  Median : 2.000   Median :11.0   Median :11.00   Median :12.00
##  Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
##  3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
```

3. Cleaning Data

```r
# 1. Merge Datasets
student_mat_data$course <- "Math"
student_por_data$course <- "Portuguese"
student_data <- bind_rows(student_mat_data, student_por_data)
```

```r
# 2. Select Columns
selected_columns <- c("sex", "age", "address", "studytime", "failures",
                      "schoolsup", "famsup", "freetime", "goout", "romantic",
                      "G1", "G2", "G3")
student_data <- student_data[, selected_columns]
```

```r
# 3. Pre-Process Categorical Data
student_data <- student_data %>%
  mutate(
    sex = ifelse(sex == "F", 0, 1),
    address = ifelse(address == "U", 0, 1),
    schoolsup = ifelse(schoolsup == "no", 0, 1),
    famsup = ifelse(famsup == "no", 0, 1),
    romantic = ifelse(romantic == "no", 0, 1)
  )
```

```r
# 4. Handle Empty Values
student_data <- na.omit(student_data)
```

```r
# 6. Display Cleared Data
head(student_data, 5)
```

```
## # A tibble: 5 x 13
##     sex   age address studytime failures schoolsup famsup freetime goout
##   <dbl> <dbl>   <dbl>     <dbl>    <dbl>     <dbl>  <dbl>    <dbl> <dbl>
## 1     0    18       0         2        0         1      0        3     4
## 2     0    17       0         2        0         0      1        3     3
## 3     0    15       0         2        3         1      0        3     2
## 4     0    15       0         3        0         0      1        2     2
## 5     0    16       0         2        0         0      1        3     2
## # i 4 more variables: romantic <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

4. Pre-Processing

```r
# 1. Functions for data preparation
student_prepare_data <- function(student_data) {
  # Split features for regression and classification
  X <- student_data %>%
    select(G1, G2, studytime, failures, freetime, goout)

  # Target for regression
  y_reg <- student_data$G3

  # Target for classification with categorization
  y_class <- cut(student_data$G3,
                 breaks = c(0, 10, 15, 20),
                 labels = c('low', 'medium', 'high'))

  # Return list with all variables
  list(X = X, y_reg = y_reg, y_class = y_class)
}
```

```r
# 2. Analysis execution
student_data_processed <- student_prepare_data(student_data)
```

4. Regression

```r
# 1. Function for regression modeling
student_perform_regression <- function(X, y_reg) {
  # Split data
  set.seed(125)
  split_index <- createDataPartition(y_reg, p = 0.8, list = FALSE)

  X_train <- X[split_index, ]
  X_test <- X[-split_index, ]
  y_train <- y_reg[split_index]
  y_test <- y_reg[-split_index]

  # Random Forest Model
  rf_model <- randomForest(
    x = X_train,
    y = y_train,
    ntree = 100,
    random_state = 125
  )

  # Predict
  y_pred <- predict(rf_model, X_test)

  # Evaluation
  mae <- mean(abs(y_test - y_pred))
  rmse <- sqrt(mean((y_test - y_pred)^2))

  # Fitur importance
```

```
  feature_importance <- data.frame(
    Feature = colnames(X),
    Importance = importance(rf_model)
  )

  # Return result
  list(
    model = rf_model,
    predictions = y_pred,
    mae = mae,
    rmse = rmse,
    feature_importance = feature_importance
  )
}
```

```
# 2. Regression
student_regression_results <- student_perform_regression(
  student_data_processed$X,
  student_data_processed$y_reg
)
```

```
# 3. Display results
cat("Regresi - Pentingnya Fitur:")
```

```
## Regresi - Pentingnya Fitur:
```

```
print(student_regression_results$feature_importance)
```

```
##             Feature IncNodePurity
## G1               G1     3760.6056
## G2               G2     6084.5543
## studytime studytime      316.5032
## failures   failures      700.0646
## freetime   freetime      282.1307
## goout         goout      333.2419
```

```
print("Regresi - Metrik:")
```

```
## [1] "Regresi - Metrik:"
```

```
print(paste("MAE:", student_regression_results$mae))
```

```
## [1] "MAE: 1.1100683318354"
```

```
print(paste("RMSE:", student_regression_results$rmse))
```

```
## [1] "RMSE: 1.79924415122392"
```

5. Classification

```r
# 1. Function for classification modeling
student_perform_classification <- function(X, y_class) {
  # Split data
  set.seed(125)
  split_index <- createDataPartition(y_class, p = 0.8, list = FALSE)

  X_train <- X[split_index, ]
  X_test <- X[-split_index, ]
  y_train <- y_class[split_index]
  y_test <- y_class[-split_index]

  # Handle Missing values
  X_train <- X_train %>%
    mutate(across(everything(), ~replace_na(., mean(., na.rm = TRUE))))

  X_test <- X_test %>%
    mutate(across(everything(), ~replace_na(., mean(., na.rm = TRUE))))

  # Decision Tree Model
  dt_model <- rpart(
    formula = y_train ~ .,
    data = data.frame(X_train, y_train),
    method = "class"
  )

  # Predict
  y_pred <- predict(dt_model, X_test, type = "class")

  # Evaluation
  conf_matrix <- confusionMatrix(y_pred, y_test)

  # Return result
  list(
    model = dt_model,
    predictions = y_pred,
    confusion_matrix = conf_matrix
  )
}
```

```r
# 2. Classification
student_classification_results <- student_perform_classification(
  student_data_processed$X,
  student_data_processed$y_class
)
```

```r
# 3. Display result
print("Klasifikasi - Confusion Matrix:")
```

```
## [1] "Klasifikasi - Confusion Matrix:"
```

```r
print(student_classification_results$confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction low medium high
##     low     64     13    0
##     medium   2     93    7
##     high     0      1   17
##
## Overall Statistics
##
##                Accuracy : 0.8832
##                  95% CI : (0.83, 0.9245)
##     No Information Rate : 0.5431
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7976
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: low Class: medium Class: high
## Sensitivity              0.9697        0.8692     0.70833
## Specificity              0.9008        0.9000     0.99422
## Pos Pred Value           0.8312        0.9118     0.94444
## Neg Pred Value           0.9833        0.8526     0.96089
## Prevalence               0.3350        0.5431     0.12183
## Detection Rate           0.3249        0.4721     0.08629
## Detection Prevalence     0.3909        0.5178     0.09137
## Balanced Accuracy        0.9352        0.8846     0.85128
```

6. Clustering

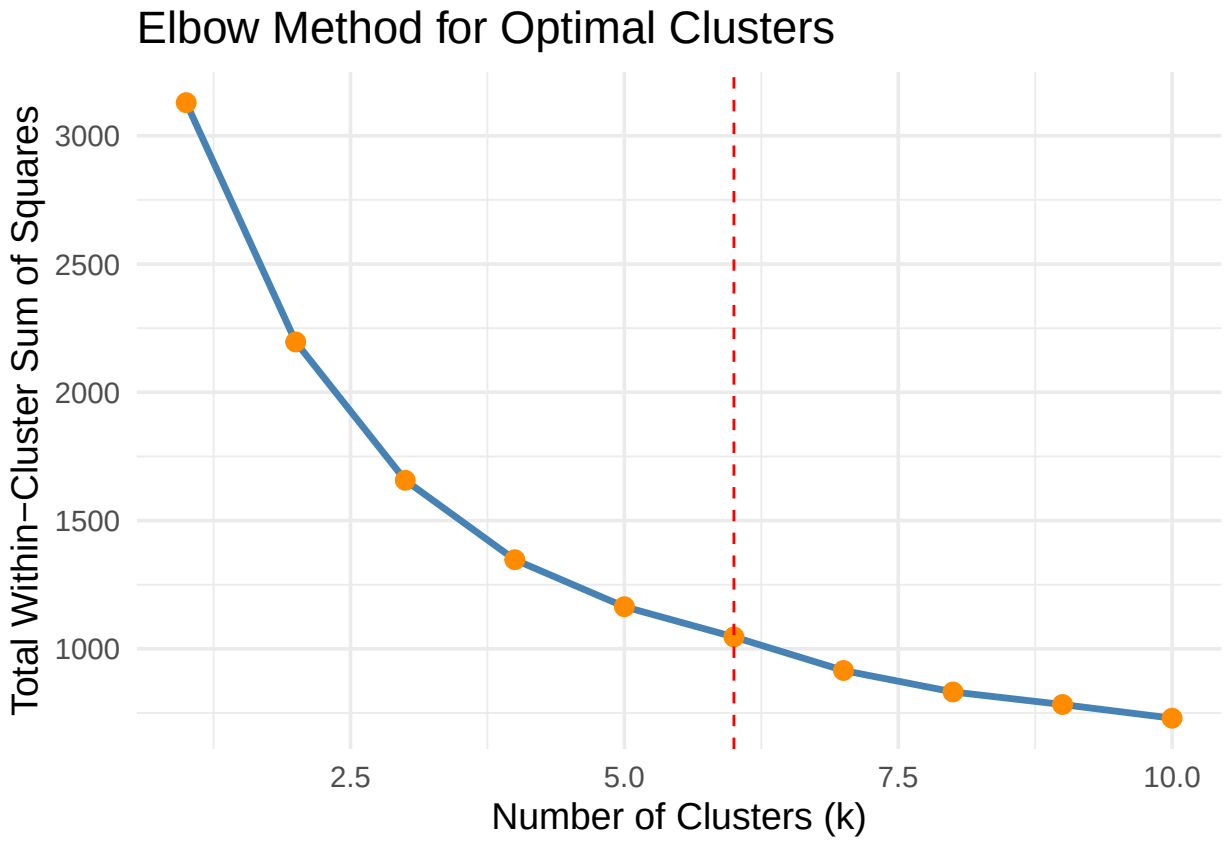
```
student_perform_clustering_analysis <- function(student_data) {
  # Prepare data for clustering
  student_clustering_data <- student_data %>%
    select(studytime, freetime, goout) %>%
    scale()

  # 1. Elbow Method with Improved Visualization
  student_elbow_method <- function(student_data) {
    wss <- sapply(1:10, function(k) {
      kmeans(student_data, centers = k, nstart = 25)$tot.withinss
    })

    optimal_k <- which(diff(diff(wss)) == min(diff(diff(wss)))) + 1  # Optimal k

    student_plt_elbow <- ggplot(data.frame(k = 1:10, wss = wss),
                                aes(x = k, y = wss)) +
      geom_line(color = "steelblue", size = 1.2) +
      geom_point(color = "darkorange", size = 3) +
```

```
    geom_vline(xintercept = optimal_k, linetype = "dashed", color = "red") +
    labs(title = "Elbow Method for Optimal Clusters",
         x = "Number of Clusters (k)",
         y = "Total Within-Cluster Sum of Squares") +
    theme_minimal(base_size = 14)

  print(student_plt_elbow)
}

student_elbow_method(student_clustering_data)

# 2. Perform K-Means clustering
student_perform_kmeans <- function(student_data, k = 3) {
  set.seed(125)
  student_km_result <- kmeans(student_data, centers = k, nstart = 25)
  student_sil <- silhouette(student_km_result$cluster, dist(student_data))

  student_plt_sil <- fviz_silhouette(student_sil, palette = "viridis") +
    labs(title = "Silhouette Plot") +
    theme_minimal(base_size = 14)

  print(student_plt_sil)

  return(list(
    student_kmeans = student_km_result,
    silhouette = student_sil
  ))
}

student_km_results <- student_perform_kmeans(student_clustering_data)

# 3. Visualize PCA with Improved Aesthetics
student_pca_visualization <- function(student_data, clusters) {
  student_pca_result <- prcomp(student_data)
  student_pca_data <- as.data.frame(student_pca_result$x[, 1:2])
  student_pca_data$Cluster <- as.factor(clusters)

  student_plt_pca <- ggplot(student_pca_data, aes(x = PC1, y = PC2, color = Cluster)) +
    geom_point(size = 3, alpha = 0.8) +
    scale_color_viridis_d() +
    geom_text(aes(label = Cluster), vjust = 2, size = 5, fontface = "bold", color = "black") +
    labs(title = "Clustering Visualization (PCA)",
         x = "Principal Component 1",
         y = "Principal Component 2") +
    theme_minimal(base_size = 14) +
    theme(legend.position = "top")

  print(student_plt_pca)
}

student_pca_visualization(student_clustering_data, student_km_results$student_kmeans$cluster)

# 4. Correlation Matrix with Gradients
```

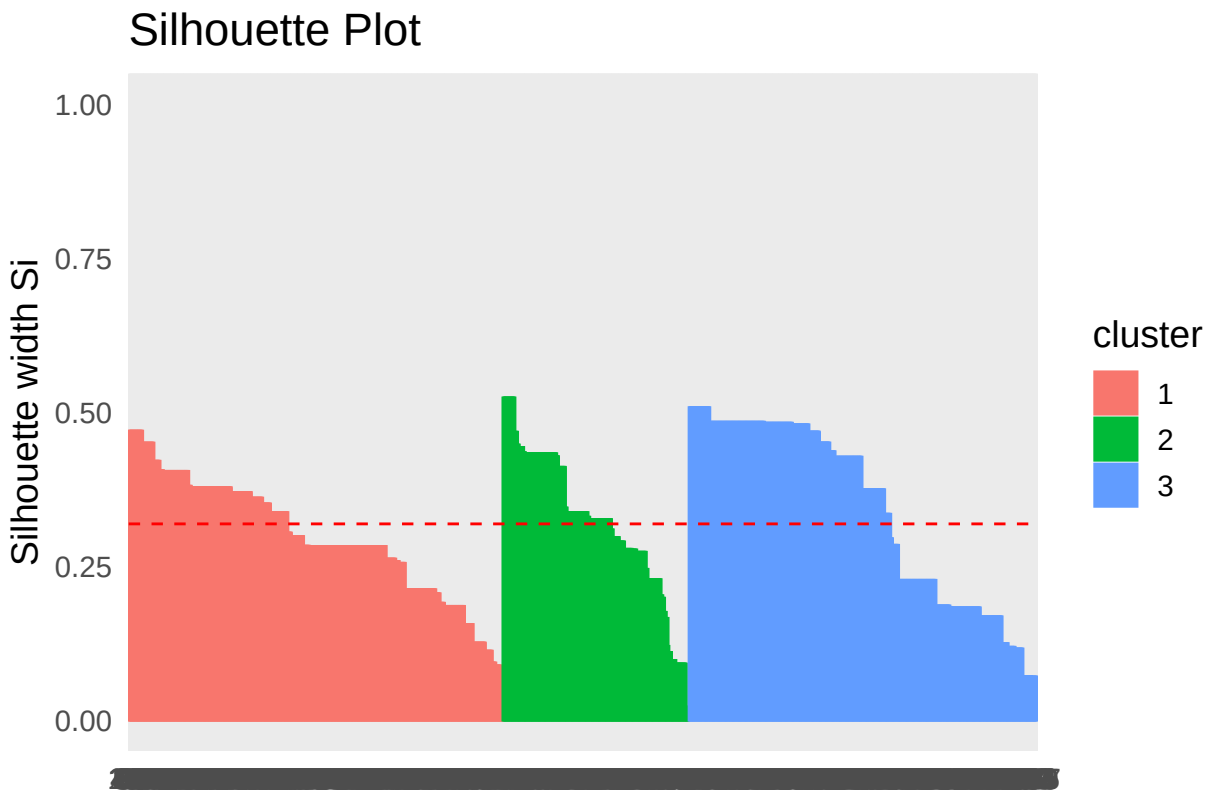
```
  student_correlation_matrix <- cor(student_data %>% select(studytime, freetime, goout, G1, G2, G3))

  corrplot(
    student_correlation_matrix,
    method = "color",
    col = viridis(10),
    type = "full",
    addCoef.col = "white",
    number.cex = 0.8,
    title = "Correlation Matrix",
    mar = c(0, 0, 2, 0)
  )

  # 5. Distribution of Performance Categories
  y_class <- cut(student_data$G2,
                 breaks = c(0, 10, 15, 20),
                 labels = c('low', 'medium', 'high'),
                 right = FALSE
                 )

  student_plt_dist <- ggplot(data.frame(y_class), aes(x = y_class)) +
    geom_bar(aes(fill = y_class), color = "black", alpha = 0.8) +
    scale_fill_viridis_d() +
    geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
    labs(title = "Distribution of Performance Categories",
         x = "Performance Category",
         y = "Number of Students") +
    theme_minimal(base_size = 14) +
    theme(legend.position = "none")

  print(student_plt_dist)

  # 6. Cluster Characteristics
  student_cluster_analysis <- student_data %>%
    mutate(Cluster = student_km_results$student_kmeans$cluster) %>%
    group_by(Cluster) %>%
    summarise(
      mean_studytime = mean(studytime),
      mean_freetime = mean(freetime),
      mean_goout = mean(goout)
    )

  print("Cluster Characteristics:")
  print(student_cluster_analysis)

  return(list(
    student_kmeans_result = student_km_results,
    student_cluster_characteristics = student_cluster_analysis
  ))
}
```

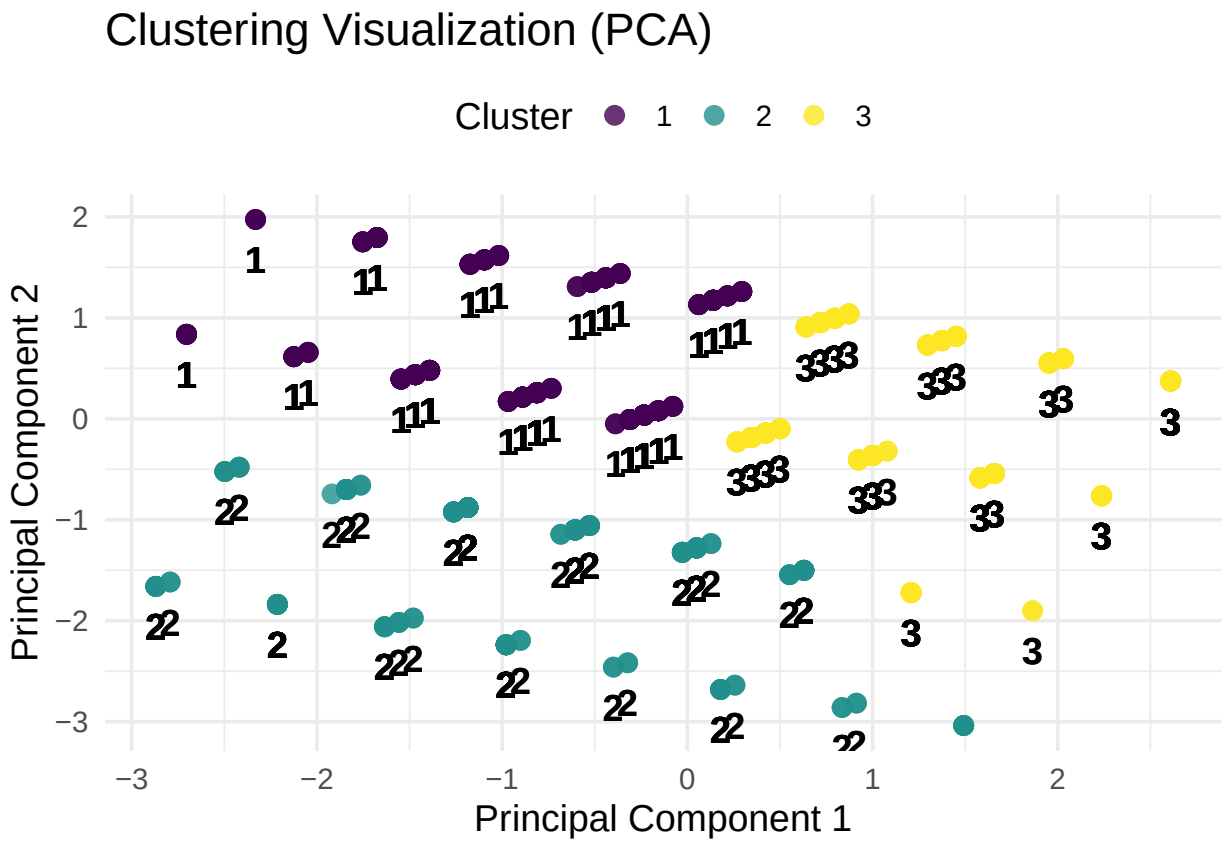
```
# Run clustering analysis
student_clustering_results <- student_perform_clustering_analysis(student_data)
```

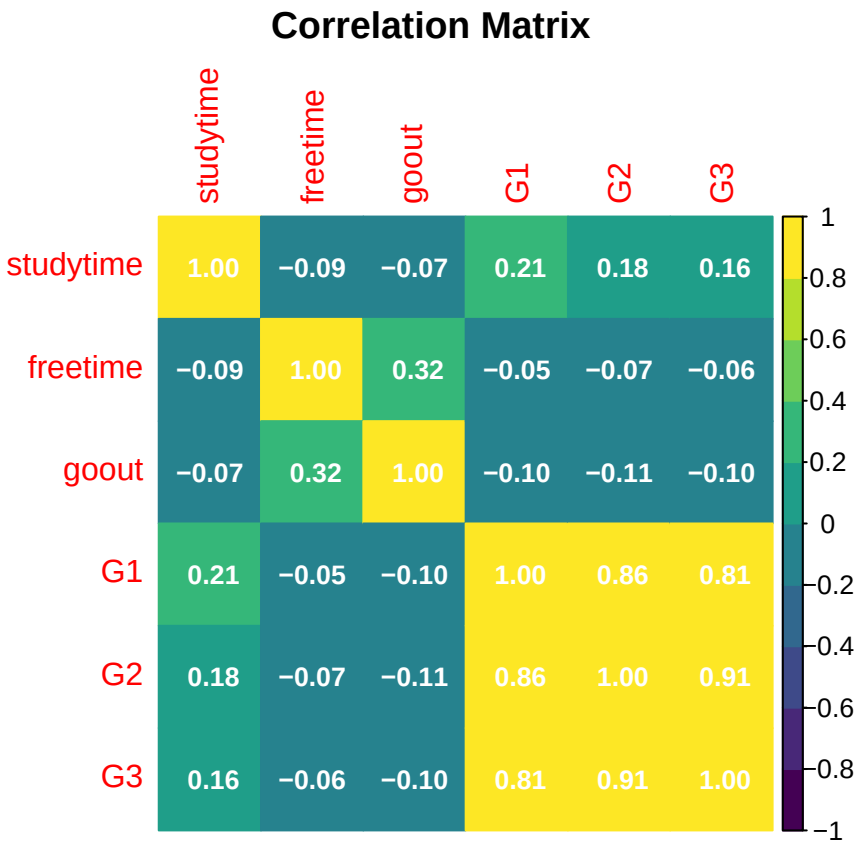
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



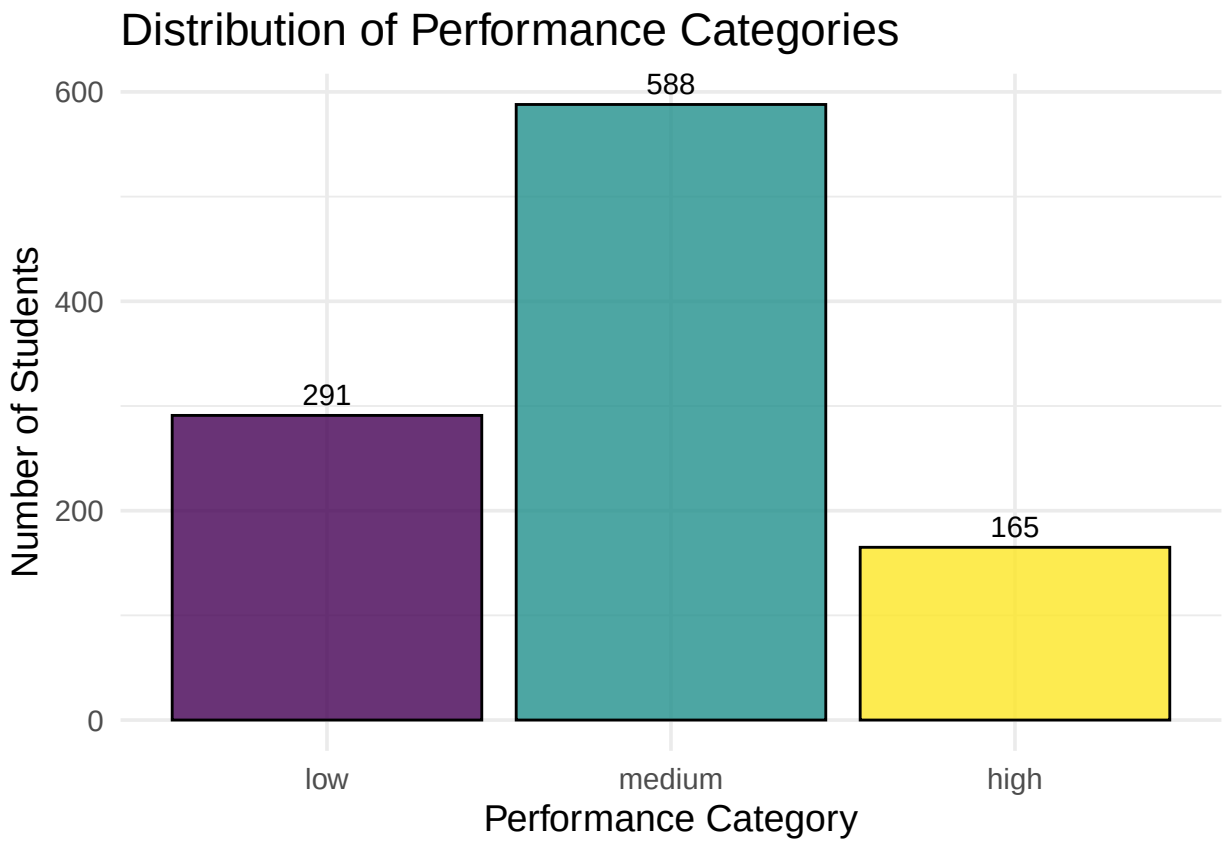

```
##   cluster size ave.sil.width
## 1       1  430          0.30
## 2       2  214          0.32
## 3       3  400          0.34
```


Silhouette Plot
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3


Clustering Visualization (PCA)
Cluster
1
2
3
Principal Component 2
Principal Component 1
2
1
0
−1
−2
−3




```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
## [1] "Cluster Characteristics:"
## # A tibble: 3 x 4
##   Cluster mean_studytime mean_freetime mean_goout
##     <int>          <dbl>         <dbl>      <dbl>
## 1       1           1.63          2.63       2.41
## 2       2           3.29          3.01       2.85
## 3       3           1.63          3.91       4.12
```