# Final Project Academic Performance Analysis

Deni Permana - Diaz Fahreza Akbar

2024-12-29

## Tahapan Exploratory Data Analysis (EDA)

### 0. Import Library dan Import Dataset

#### Import Library

Memuat paket yang diperlukan

```
if (!requireNamespace("readr")) install.packages("readr")
```

```
## Loading required namespace: readr
```

```
if (!requireNamespace("dplyr")) install.packages("dplyr")
```

```
## Loading required namespace: dplyr
```

```
if (!requireNamespace("ggplot2")) install.packages("ggplot2")
```

```
## Loading required namespace: ggplot2
```

```
if (!requireNamespace("gridExtra")) install.packages("gridExtra")
```

```
## Loading required namespace: gridExtra
```

```
if (!requireNamespace("reshape2")) install.packages("reshape2")
```

```
## Loading required namespace: reshape2
```

```
library(readr)  # Untuk membaca data dengan cepat dari berbagai format seperti CSV
library(dplyr)  # Untuk manipulasi data dengan sintaks yang bersih dan intuitif
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)  # Paket grafis yang powerful untuk membuat visualisasi data
library(gridExtra)  # Untuk menampilkan beberapa plot dalam satu tampilan
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(reshape2)  # Memudahkan mengubah data antara format lebar dan panjang
```

**Penjelasan:** Memuat paket(library) yang diperlukan. 'readr' untuk membaca data dari file CSV; 'dplyr' untuk manipulasi data; 'ggplot2' untuk visualisasi data; 'gridExtra' menampikan beberapa plot; dan 'reshape2' untuk mengubah format data

### Import Dataset

Membaca dataset dari file CSV ke dalam R.

```
# 1. Datset 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.
```

```
# 2. Dataset 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.
```

**Penjelasan:** Membaca dua dataset menggunakan fungsi read_delim() dari library readr. Dataset pertama, student-mat.csv, memuat data siswa terkait mata pelajaran matematika, sementara dataset kedua, student-por.csv, berisi data siswa untuk mata pelajaran bahasa Portugis. Kedua file tersebut menggunakan titik koma (;) sebagai pembatas, dan fungsi ini memastikan spasi tambahan di awal atau akhir nilai dihapus untuk menjaga kebersihan data. Output dari masing-masing proses adalah data frame (student_mat_data dan student_por_data)

## 1. Data Understanding

### 1.1 Tujuan:

Memahami konteks, struktur, dan karakteristik dataset. Ini mencakup identifikasi variabel, analisis atribut, dan pemahaman tentang nilai yang hilang.

### 1.2 Hasil yang Diharapkan:

Gambaran umum tentang dataset, termasuk dimensi, tipe data, dan potensi isu yang perlu diatasi sebelum analisis lebih lanjut.

### 1.3 Langkah-langkah:

#### 1.3.1 Deskripsi Dataset:

| No | Nama Variabel | Type | Deskripsi |
|---|---|---|---|
| 1 | school | Categorical | ‘student’s school (GP’ - Gabriel Pereira atau ‘MS’ - Mousinho da Silveira) |
| 2 | sex | Binary | student’s sex (binary: ‘F’ - female or ‘M’ - male) |
| 3 | age | Integer | student’s age (numeric: from 15 to 22) |
| 4 | addres | Categorical | student’s home address type (binary: ‘U’ - urban or ‘R’ - rural) |
| 5 | famsize | Categorical | family size (binary: ‘LE3’ - less or equal to 3 or ‘GT3’ - greater than 3) |
| 6 | Pstatus | Categorical | parent’s cohabitation status (binary: ‘T’ - living together or ‘A’ - apart) |
| 7 | Medu | Integer | mother’s education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) |
| 8 | Fedu | Integer | father’s education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education) |
| 9 | Mjob | Categorical | mother’s job (nominal: ‘teacher’, ‘health’ care related, civil ‘services’ (e.g. administrative or police), ‘at_home’ or ‘other’) |
| 10 | Fjob | Categorical | father’s job (nominal: ‘teacher’, ‘health’ care related, civil ‘services’ (e.g. administrative or police), ‘at_home’ or ‘other’) |
| 11 | reason | Categorical | reason to choose this school (nominal: close to ‘home’, school ‘reputation’, ‘course’ preference or ‘other’) |
| 12 | guardian | Categorical | reason to choose this school (nominal: close to ‘home’, school ‘reputation’, ‘course’ preference or ‘other’) |

| No | Nama Variabel | Type | Deskripsi |
|---|---|---|---|
| 13 | traveltime | Integer | home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| 14 | studytime | Integer | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| 15 | failures | Integer | number of past class failures (numeric: n if 1<=n<3, else 4) |
| 16 | schoolsup | Binary | extra educational support (binary: yes or no) |
| 17 | famsup | Binary | family educational support (binary: yes or no) |
| 18 | paid | Binary | extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) |
| 19 | activities | Binary | extra-curricular activities (binary: yes or no) |
| 20 | nursery | Binary | attended nursery school (binary: yes or no) |
| 21 | higher | Binary | wants to take higher education (binary: yes or no) |
| 22 | internet | Binary | Internet access at home (binary: yes or no) |
| 23 | romantic | Binary | with a romantic relationship (binary: yes or no) |
| 24 | famrel | Integer | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| 25 | freetime | Integer | free time after school (numeric: from 1 - very low to 5 - very high) |
| 26 | goout | Integer | going out with friends (numeric: from 1 - very low to 5 - very high) |
| 27 | Dalc | Integer | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 28 | Walc | Integer | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 29 | health | Integer | current health status (numeric: from 1 - very bad to 5 - very good) |
| 30 | absences | Integer | number of school absences (numeric: from 0 to 93) |
| 31 | G1 | Categorical | first period grade (numeric: from 0 to 20) |
| 32 | G2 | Categorical | second period grade (numeric: from 0 to 20) |
| 33 | G3 | Integer | final grade (numeric: from 0 to 20, output target) |

```
cat(dim(student_mat_data), "\n")  # Menampilkan dimensi dataset mat
```

#### 1.3.1.1 Menghitung jumlah baris dan kolom

```
## 395 33
```

**Penjelasan:** Untuk dataset mat memiliki 395 baris data dan 33 kolom yang berbeda

```
cat(dim(student_por_data), "\n")  # Menampilkan dimensi dataset por
```

```
## 649 33
```

**Penjelasan:** Untuk dataset por memiliki 649 baris data dan 33 kolom yang berbeda

```r
print(head(student_mat_data, 5))  # Menampilkan 5 baris pertama dataset mat
```

#### 1.3.1.2 Tampilkan beberapa baris pertama dari masing-masing dataset

```
## # 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>
```

**Penjelasan:** Menampilkan 5 baris data pada dataset mat dengan kolom pertama atau paling kiri adalah 'school' dan kolom terakhir atau paling kanan adalah 'G3'

```r
print(head(student_por_data, 5))  # Menampilkan 5 baris pertama dataset por
```

```
## # 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>
```

**Penjelasan:** Menampilkan 5 baris data pada dataset por dengan kolom pertama atau paling kiri adalah 'school' yang berisi data 'GP' dan kolom terakhir atau paling kanan adalah 'G3'

**1.3.1.3 Analisis Struktur Data** Menggunakan str() untuk melihat struktur dan tipe data dari setiap kolom.

```r
str(student_mat_data)  # Menampilkan struktur dataset mat
```

```
## 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>
```

**Penjelasan:** Output dari kode str(student_mat_data) menampilkan struktur dataset student_mat_data yang terdiri dari 395 baris dan 33 kolom. Setiap kolom memiliki tipe data yang berbeda, seperti karakter untuk kolom school, sex, address, dan lainnya, serta numerik untuk kolom age, Medu, Fedu, dan seterusnya. Informasi ini membantu dalam memahami struktur data sebelum melakukan analisis lebih lanjut.

```r
str(student_por_data)  # Menampilkan struktur dataset por
```

```
## 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>
```

**Penjelasan:** Output dari kode str(student_por_data) menampilkan struktur dataset student_por_data yang terdiri dari 649 baris dan 33 kolom. Sama seperti student_mat_data, setiap kolom dalam dataset ini juga memiliki tipe data yang bervariasi, dengan kolom karakter seperti school, sex, address, dan kolom numerik seperti age, Medu, Fedu, dan lainnya. Informasi ini memberikan gambaran awal tentang data yang akan dianalisis.

**1.3.2 Identifikasi Variabel:** Menampilkan nama variabel yang terdapat dalam masing-masing dataset.

```r
print(names(student_mat_data))  # Menampilkan nama kolom dataset mat
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
```

```
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "G3"
```

**Penjelasan:** Menampilkan seluruh nama kolom yang ada pada dataset mat yang dimulai dari 'school' untuk kolom pertama dan 'G3' untuk kolom terakhir

```
print(names(student_por_data))  # Menampilkan nama kolom dataset por
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "G3"
```

**Penjelasan:** Menampilkan seluruh nama kolom yang ada pada dataset mat yang dimulai dari 'school' untuk kolom pertama dan 'G3' untuk kolom terakhir

**1.3.3 Analisis Missing Values:** Menghitung dan menampilkan jumlah nilai yang hilang dalam setiap kolom.

```
print(colSums(is.na(student_mat_data)))  # Menampilkan jumlah missing values per kolom dataset mat
```

```
##     school        sex        age    address    famsize    Pstatus       Medu
##          0          0          0          0          0          0          0
##       Fedu       Mjob       Fjob     reason   guardian traveltime  studytime
##          0          0          0          0          0          0          0
##   failures  schoolsup     famsup       paid activities    nursery     higher
##          0          0          0          0          0          0          0
##   internet   romantic     famrel   freetime      goout       Dalc       Walc
##          0          0          0          0          0          0          0
##     health   absences         G1         G2         G3
##          0          0          0          0          0
```

**Penjelasan:** Semua kolom ditampilkan kembali tetapi dengan menghitung jumlah nilai yang kosong dari baris data mat yang ada 395 baris dan pada tiap kolom tidak ditemukan nilai yang kosong yang dilihat dari angka dibawah kolomnya 0 (nol).

```
print(colSums(is.na(student_por_data)))  # Menampilkan jumlah missing values per kolom dataset por
```

```
##     school        sex        age    address    famsize    Pstatus       Medu
##          0          0          0          0          0          0          0
##       Fedu       Mjob       Fjob     reason   guardian traveltime  studytime
##          0          0          0          0          0          0          0
##   failures  schoolsup     famsup       paid activities    nursery     higher
##          0          0          0          0          0          0          0
##   internet   romantic     famrel   freetime      goout       Dalc       Walc
##          0          0          0          0          0          0          0
##     health   absences         G1         G2         G3
##          0          0          0          0          0
```

**Penjelasan:** Semua kolom ditampilkan kembali tetapi dengan menghitung jumlah nilai yang kosong dari baris data mat yang ada 649 baris dan pada tiap kolom tidak ditemukan nilai yang kosong yang dilihat dari angka dibawah kolomnya 0 (nol).

## 2. Exploratory Data Analysis Awal

### 2.1 Tujuan:

Melihat data secara umum dan mendapatkan insight awal. Ini termasuk penggunaan statistik deskriptif dan visualisasi untuk memahami distribusi dan sebaran data.

### 2.2 Hasil yang Diharapkan:

Pemahaman mendalam tentang distribusi nilai, outlier, dan pola awal yang dapat diidentifikasi dalam data.

### 2.3 Langkah-langkah:

```
print(summary(student_mat_data))  # Menampilkan ringkasan statistik dataset mat
```

#### 2.3.1 Statistik Deskriptif:

```
##     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
```

**Penjelasan:** Ringkasan ini mencakup informasi seperti jumlah baris, tipe data, dan statistik dasar untuk setiap kolom. Misalnya, untuk kolom age, ringkasan ini menunjukkan nilai minimum, kuartil pertama, median, mean, kuartil ketiga, dan nilai maksimum. Selain itu, ringkasan ini juga mencakup informasi tentang distribusi nilai untuk kolom-kolom numerik lainnya seperti Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, dan G3. Untuk kolom karakter seperti school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, dan romantic, ringkasan ini memberikan panjang dan tipe data dari setiap kolom.

```r
print(summary(student_por_data))  # Menampilkan ringkasan statistik dataset por
```

```
##     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
```

**Penjelasan:** Seperti pada dataset student_mat_data, ringkasan ini mencakup informasi dasar seperti jumlah baris, tipe data, dan statistik deskriptif untuk setiap kolom. Untuk kolom numerik seperti age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, dan G3, ringkasan ini menunjukkan nilai minimum, kuartil pertama, median, mean, kuartil ketiga, dan nilai maksimum. Sedangkan untuk kolom karakter seperti school, sex, address, famsize, Pstatus, Mjob, Fjob,

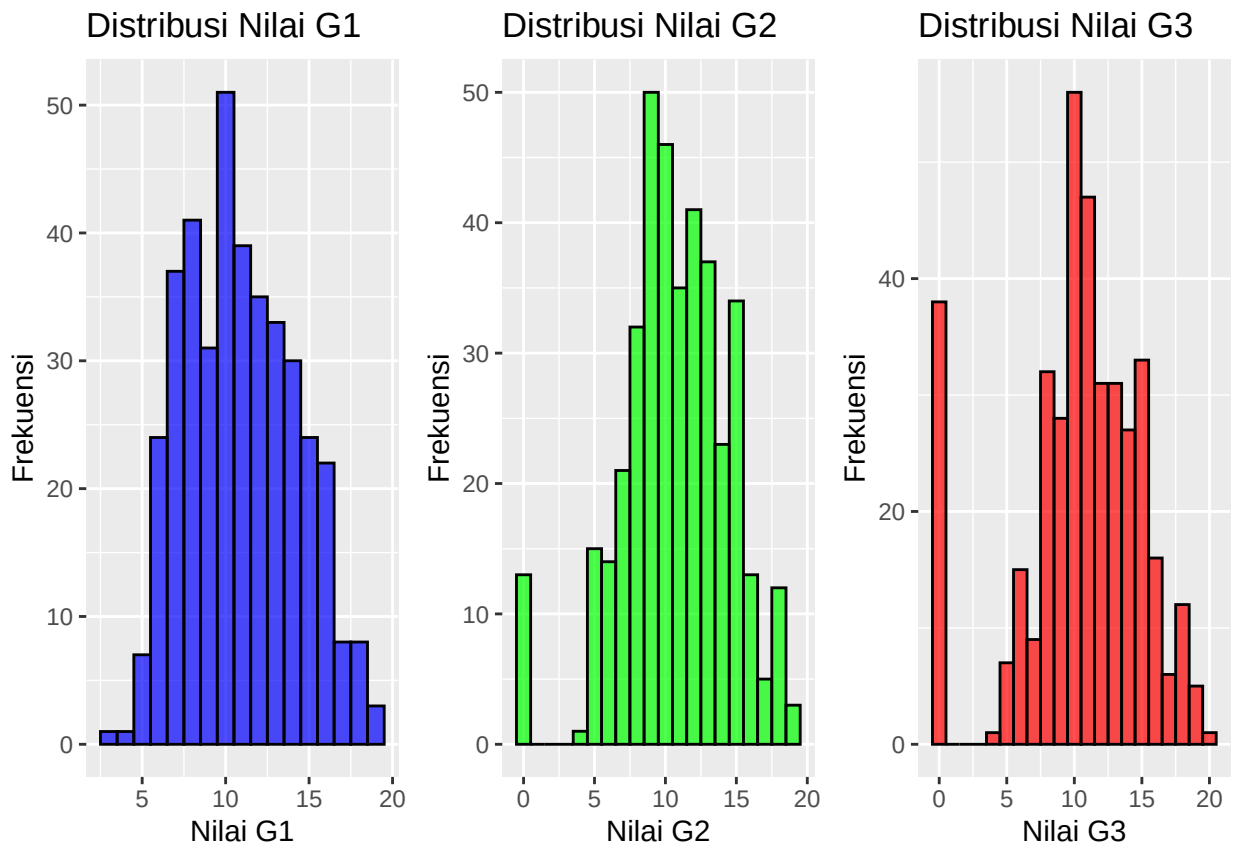
reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, dan romantic, ringkasan ini memberikan panjang dan tipe data dari setiap kolom. Ringkasan ini membantu dalam memahami distribusi dan karakteristik data sebelum melakukan analisis lebih lanjut.

### 2.3.2 Visualisasi Awal:

**2.3.2.1 Histogram:** Histogram untuk Nilai G1, G2, G3

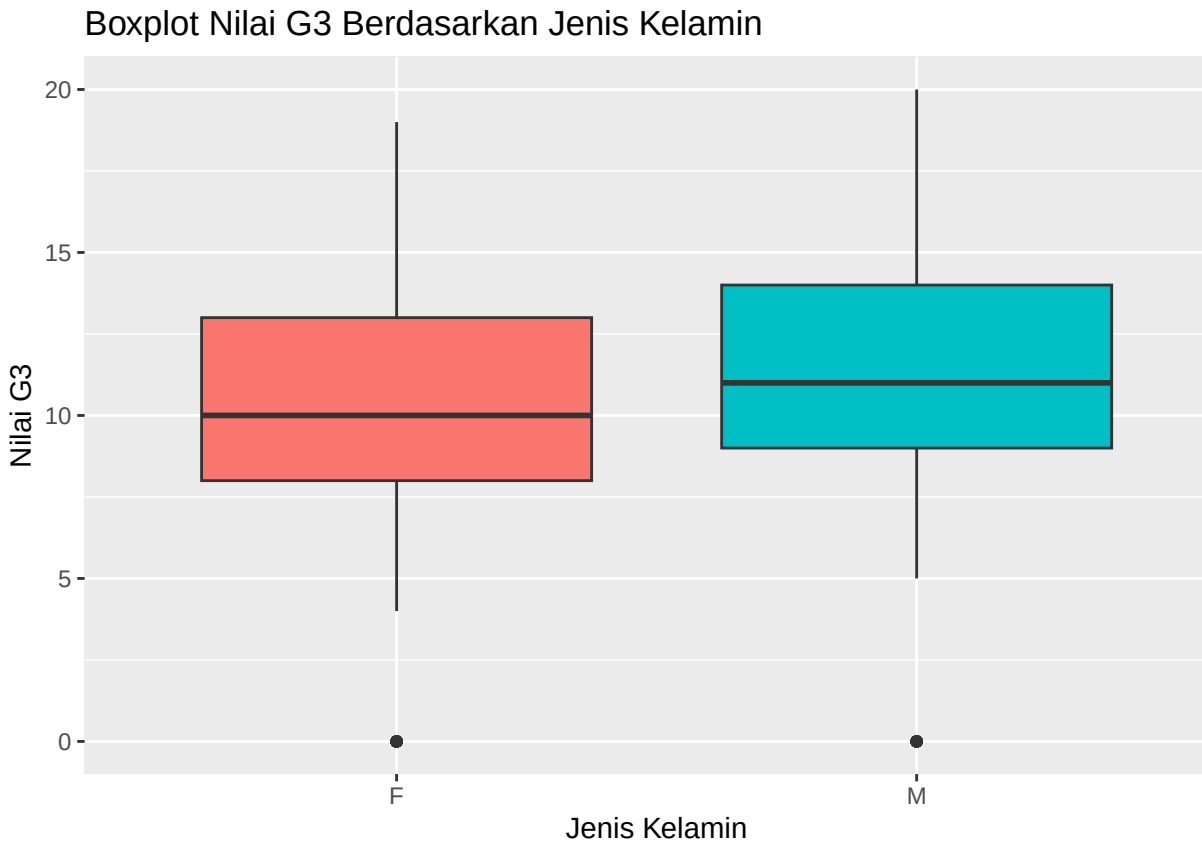
```
# G1
g1_hist <- ggplot(student_mat_data, aes(x = G1)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribusi Nilai G1", x = "Nilai G1", y = "Frekuensi")

# G2
g2_hist <- ggplot(student_mat_data, aes(x = G2)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black", alpha = 0.7) +
  labs(title = "Distribusi Nilai G2", x = "Nilai G2", y = "Frekuensi")

# G3
g3_hist <- ggplot(student_mat_data, aes(x = G3)) +
  geom_histogram(binwidth = 1, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Distribusi Nilai G3", x = "Nilai G3", y = "Frekuensi")
```

```
# Tampilkan histogram dalam satu tampilan
grid.arrange(g1_hist, g2_hist, g3_hist, ncol = 3)
```



**Penjelasan:** Histogram menunjukkan distribusi nilai G1, G2, dan G3. Dari histogram ini, kita dapat melihat bahwa sebagian besar siswa memiliki nilai di kisaran menengah, dengan beberapa siswa yang memiliki nilai sangat tinggi atau sangat rendah.

```
# Boxplot untuk Nilai G3 Berdasarkan Jenis Kelamin
boxplot_gender <- ggplot(student_mat_data, aes(x = sex, y = G3, fill = factor(sex))) +
  geom_boxplot() +
  labs(title = "Boxplot Nilai G3 Berdasarkan Jenis Kelamin",
       x = "Jenis Kelamin", y = "Nilai G3") +
  scale_x_discrete(labels = c("0" = "Female", "1" = "Male")) +
  theme(legend.position = "none")

print(boxplot_gender)
```

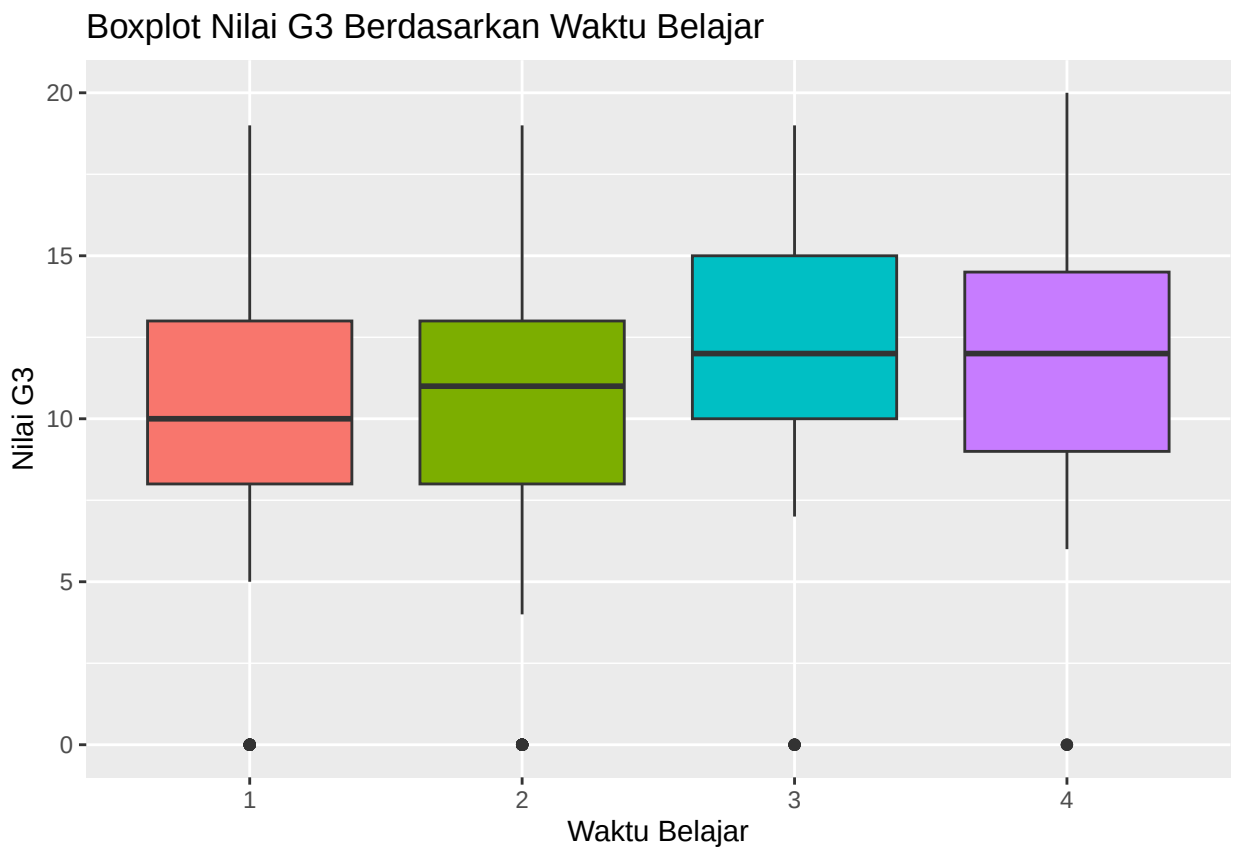



#### 2.3.2.2 Boxplot:

Boxplot ini menunjukkan bahwa tidak ada perbedaan signifikan antara nilai G3 siswa laki-laki dan perempuan. Kedua kelompok memiliki distribusi nilai yang serupa.

```
# Boxplot untuk Nilai G3 Berdasarkan Waktu Belajar
boxplot_studytime <- ggplot(student_mat_data, aes(x = factor(studytime), y = G3, fill = factor(studytime
  geom_boxplot() +
  labs(title = "Boxplot Nilai G3 Berdasarkan Waktu Belajar",
       x = "Waktu Belajar", y = "Nilai G3") +
  theme(legend.position = "none")
```

```r
print(boxplot_studytime)
```




**Penjelasan:** Boxplot ini menunjukkan bahwa siswa yang menghabiskan lebih banyak waktu untuk belajar cenderung memiliki nilai G3 yang lebih tinggi. Ini menunjukkan adanya hubungan positif antara waktu belajar dan nilai akhir.

```r
# Analisis Outlier menggunakan Boxplot
outliers <- boxplot(student_mat_data$G3, plot = FALSE)$out
print(outliers)  # Menampilkan nilai outlier
```

```
## numeric(0)
```

**Penjelasan:** Menampilkan nilai outlier, dikarenakan tidak ada outlier maka nilanya 0(nol)

## 3. Visualisasi Data

### 3.1 Tujuan:

Membuat visualisasi yang menarik untuk menggambarkan hubungan antar variabel dan pola dalam dataset. Visualisasi ini membantu dalam menganalisis dan menceritakan data secara lebih efektif.

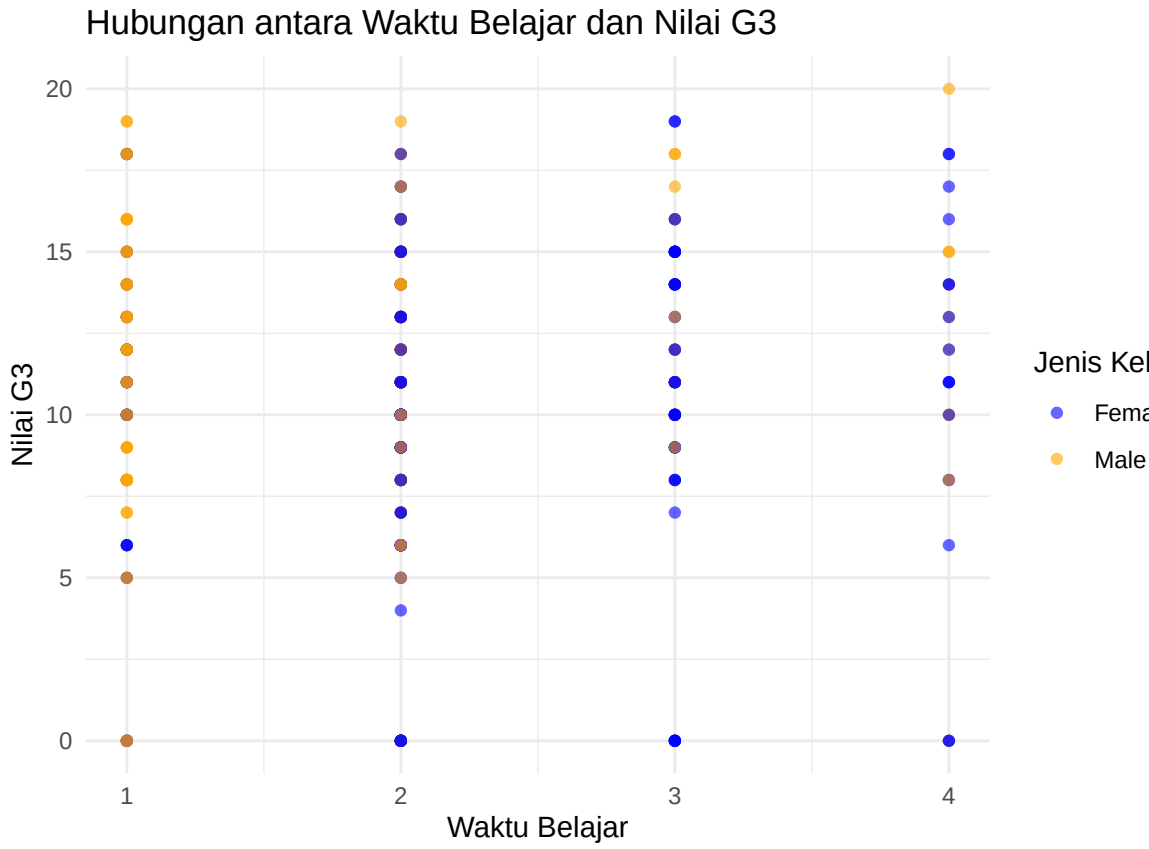
### 3.2 Hasil yang Diharapkan:

Visualisasi yang jelas dan informatif yang menunjukkan hubungan, tren, dan pola dalam data, serta mendukung narasi analisis.

### 3.3 Langkah-langkah:

#### 3.3.1 Hubungan antar Variabel:

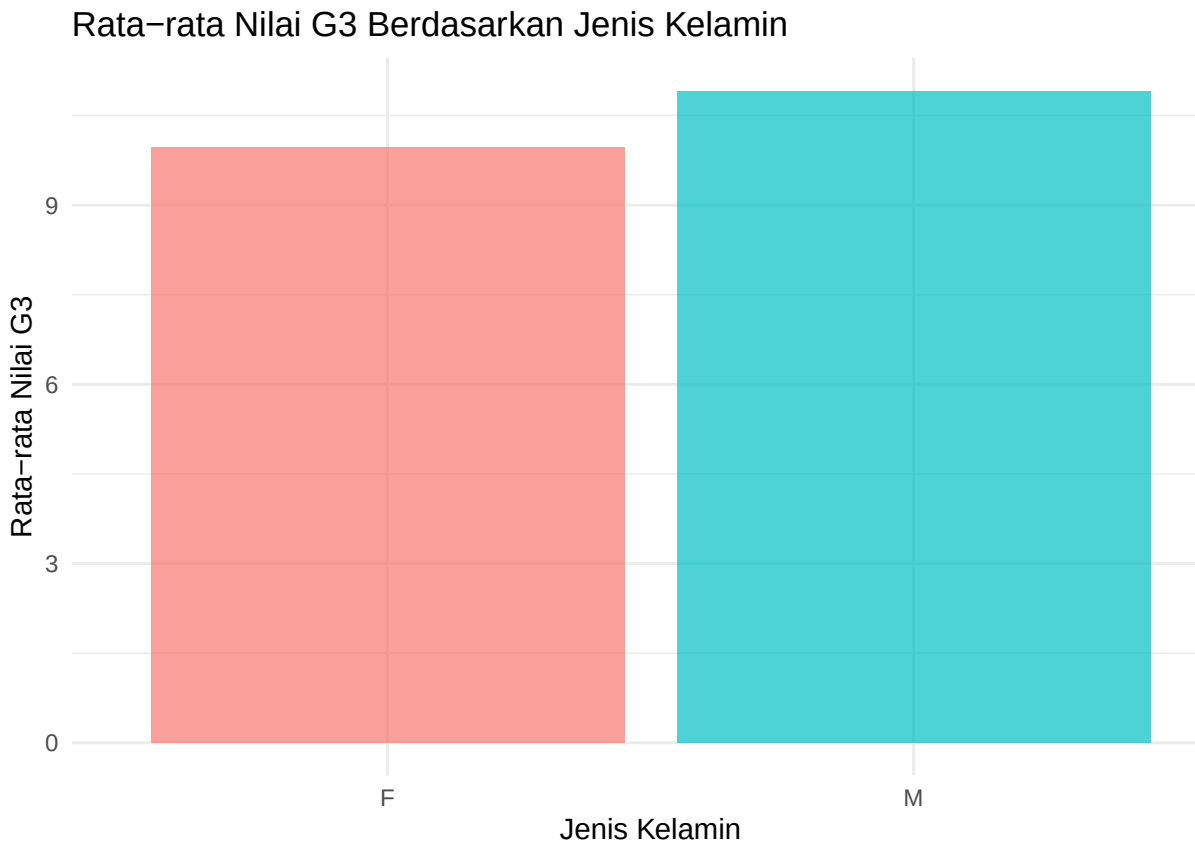
```
# Scatter Plot: Hubungan antara Waktu Belajar dan Nilai G3
scatter_studytime_g3 <- ggplot(student_mat_data, aes(x = studytime, y = G3, color = factor(sex))) +
  geom_point(alpha = 0.6) +
  labs(title = "Hubungan antara Waktu Belajar dan Nilai G3",
       x = "Waktu Belajar",
       y = "Nilai G3",
       color = "Jenis Kelamin") +
  scale_color_manual(values = c("blue", "orange"), labels = c("Female", "Male")) +
  theme_minimal()

print(scatter_studytime_g3)
```




##### 3.3.1.1 Scatter Plot:

**Penjelasan:** Scatter plot ini memperkuat temuan dari boxplot sebelumnya, menunjukkan bahwa ada tren positif antara waktu belajar dan nilai G3. Siswa yang menghabiskan lebih banyak waktu untuk belajar cenderung mendapatkan nilai yang lebih baik.

```r
# Bar Plot: Rata-rata Nilai G3 Berdasarkan Jenis Kelamin
avg_g3_gender <- student_mat_data %>%
  group_by(sex) %>%
  summarise(mean_G3 = mean(G3, na.rm = TRUE))

barplot_gender <- ggplot(avg_g3_gender, aes(x = factor(sex), y = mean_G3, fill = factor(sex))) +
  geom_bar(stat = "identity", alpha = 0.7) +
  labs(title = "Rata-rata Nilai G3 Berdasarkan Jenis Kelamin",
       x = "Jenis Kelamin",
       y = "Rata-rata Nilai G3") +
  scale_x_discrete(labels = c("0" = "Female", "1" = "Male")) +
  theme_minimal() +
  theme(legend.position = "none")

print(barplot_gender)
```

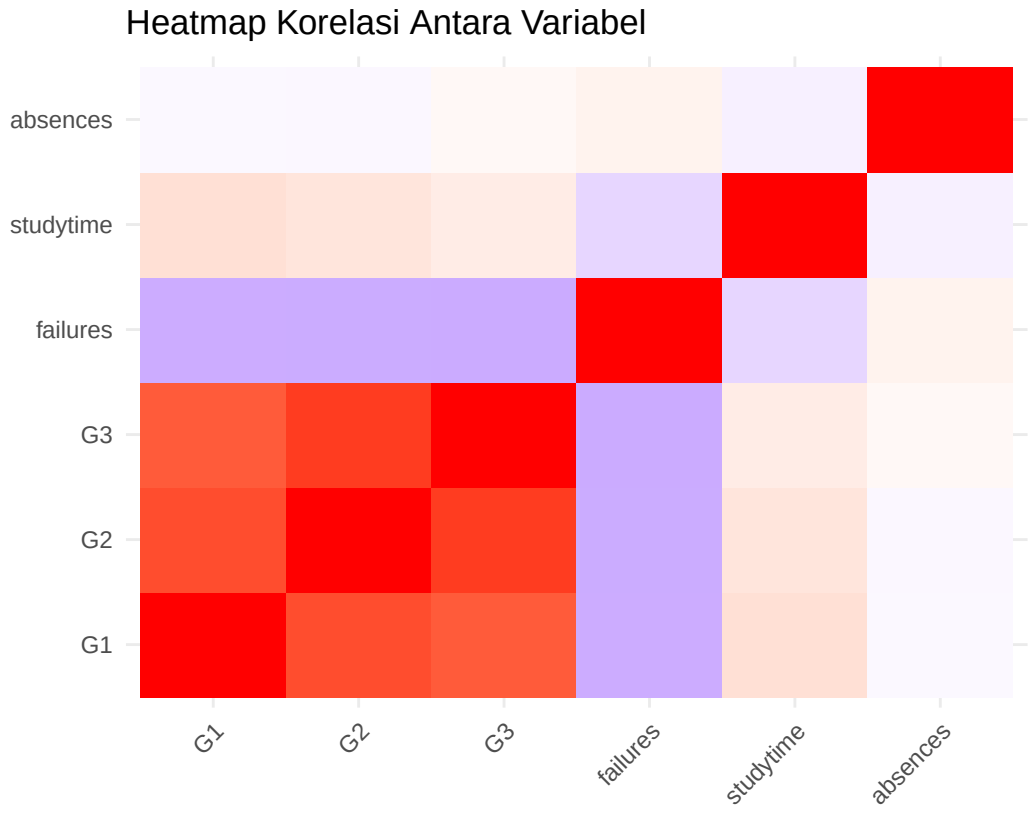



#### 3.3.1.2 Bar Plot:

**Penjelasan:** Bar plot ini menunjukkan bahwa rata-rata nilai G3 untuk siswa laki-laki dan perempuan hampir sama, mengindikasikan bahwa jenis kelamin tidak memiliki pengaruh signifikan terhadap nilai akhir.

```r
# Heatmap: Korelasi antar Variabel Numerik
correlation_matrix <- cor(student_mat_data %>% select(G1, G2, G3, failures, studytime, absences), use =
```

```r
## Mengubah matriks korelasi menjadi format yang bisa digunakan untuk ggplot
correlation_melted <- melt(correlation_matrix)

heatmap_plot <- ggplot(correlation_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name="Korelasi") +
  labs(title = "Heatmap Korelasi Antara Variabel", x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(heatmap_plot)
```




### 3.3.2 Heatmap/Korelasi:

**Penjelasan:** Heatmap ini menunjukkan korelasi antara variabel-variabel numerik dalam dataset. Misalnya, terdapat korelasi positif yang kuat antara nilai G1, G2, dan G3, yang menunjukkan bahwa siswa yang mendapatkan nilai baik di awal cenderung mempertahankan performa mereka.

```r
# Pie Chart: Proporsi Dukungan Keluarga
family_support <- student_mat_data %>%
  group_by(famsup) %>%
  summarise(count = n())
```

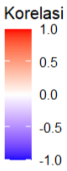
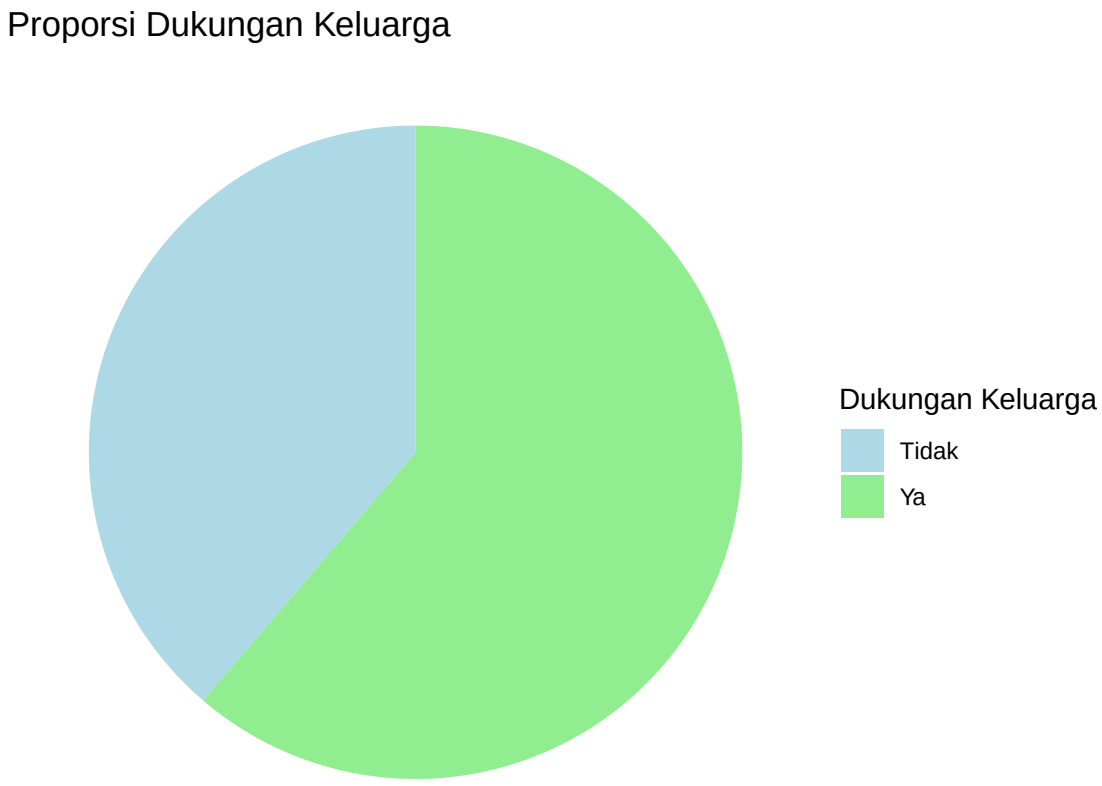



Figure 1: nilai 1.0 menunjukan korelasi positif yang kuat, nilai -1.0 menunjukan korelasi negatif yang kuat, dan jika nilai mendekati nol artinya tidak ada korelasi yang jelas. jadi semakin nilainya mendekati nilai positif maka tren nya positif begitupun sebaliknya.

```
pie_chart <- ggplot(family_support, aes(x = "", y = count, fill = factor(famsup))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Proporsi Dukungan Keluarga",
       fill = "Dukungan Keluarga") +
  scale_fill_manual(values = c("lightblue", "lightgreen"), labels = c("Tidak", "Ya")) +
  theme_void()

print(pie_chart)
```




### 3.3.3 Pie Chart:

Pie chart ini menunjukkan bahwa sebagian besar siswa menerima dukungan keluarga. Dukungan keluarga dapat menjadi faktor penting dalam keberhasilan akademik siswa.

## 4. Kesimpulan dan Rekomendasi

### 4.1 Kesimpulan Utama:

- Gender: Tidak ada perbedaan signifikan dalam nilai berdasarkan gender.
- Waktu Belajar: Waktu belajar yang lebih banyak berkorelasi dengan nilai yang lebih tinggi.
- Dukungan Keluarga: Sebagian besar siswa menerima dukungan keluarga, yang penting untuk keberhasilan akademik.
- Korelasi Nilai: Nilai awal yang baik berkorelasi dengan nilai akhir yang baik.

### 4.2 Rekomendasi:

- Intervensi Pendidikan: Fokus pada siswa yang memiliki nilai awal rendah untuk memberikan dukungan tambahan.
- Dukungan Keluarga: Mendorong partisipasi aktif orang tua dalam pendidikan anak mereka.
- Alokasi Waktu Belajar: Membantu siswa mengembangkan strategi belajar yang efektif dan manajemen waktu.
- Keadilan Gender: Terus memantau dan memastikan bahwa tidak ada bias gender dalam evaluasi dan pendidikan.

# Tahapan Data Analysis

## 0. Import Library

```
if (!requireNamespace("tidyverse")) install.packages("tidyverse")
```

```
## Loading required namespace: tidyverse
```

```
if (!requireNamespace("caret")) install.packages("caret")
```

```
## Loading required namespace: caret
```

```
if (!requireNamespace("randomForest")) install.packages("randomForest")
```

```
## Loading required namespace: randomForest
```

```
if (!requireNamespace("rpart")) install.packages("rpart")
if (!requireNamespace("cluster")) install.packages("cluster")
```

```
## Loading required namespace: cluster
```

```
if (!requireNamespace("factoextra")) install.packages("factoextra")
```

```
## Loading required namespace: factoextra
```

```r
if (!requireNamespace("corrplot")) install.packages("corrplot")
```

```
## Loading required namespace: corrplot
```

```r
if (!requireNamespace("stats")) install.packages("stats")
if (!requireNamespace("NbClust")) install.packages("NbClust")
```

```
## Loading required namespace: NbClust
```

```r
if (!requireNamespace("viridis")) install.packages("viridis")
```

```
## Loading required namespace: viridis
```

```r
library(tidyverse)  # Sekumpulan paket terintegrasi untuk analisis data yang mendukung tidy data
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## 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
```

```r
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
```

```r
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:gridExtra':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
```

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

```r
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
```

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

```
## corrplot 0.95 loaded
```

```r
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
```

**Penjelasan:** Memuat tambahan paket yang diperlukan untuk analisis data.

## 1. 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)
```

**Penjelasan:** Kolom baru ditambahkan untuk menunjukkan jenis kursus (matematika atau bahasa Portugis) dalam masing-masing dataset. Kemudian, kedua dataset digabung menggunakan bind_rows() untuk membuat satu dataset komprehensif (student_data) yang mencakup semua informasi siswa.

```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]
```

**Penjelasan:** Memilih kolom-kolom yang relevan untuk analisis lebih lanjut. Dengan menyimpan hanya kolom yang diperlukan(kolom yang dipilih: sex, age, address, studytime, failures, schoolsup, famsup, freetime, goout, romantic, G1, G2, G3 ), dataset menjadi lebih ringkas dan fokus pada variabel yang berdampak pada performa akademik.

```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)
  )
```

**Penjelasan:** Data kategorikal diubah menjadi format numerik untuk mempermudah analisis. Misalnya, jenis kelamin (sex) diubah menjadi 0 (perempuan) dan 1 (laki-laki), serta variabel lainnya seperti alamat, dukungan sekolah, dukungan keluarga, dan status hubungan juga dikonversi ke 0 dan 1. Ini membuat data lebih siap untuk digunakan dalam model analisis.

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

**Penjelasan:** Menghapus semua baris yang memiliki nilai kosong (NA) dari dataset.

```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>
```

**Penjelasan:** Menampilkan lima baris pertama dari dataset yang telah dibersihkan.

## 2. 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)
}
```

**Penjelasan:** Data siswa dipersiapkan dengan cara yang sistematis untuk memfasilitasi analisis lebih lanjut. Dengan memilih fitur yang relevan dan mendefinisikan target untuk model (regresi dan klasifikasi), fungsi ini menetapkan dasar yang kuat untuk langkah-langkah pemodelan yang akan datang.

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

**Penjelasan:** Mengimplementasikan langkah persiapan data yang telah didefinisikan sebelumnya. Dengan memproses data siswa dan menyimpannya dalam variabel student_data_processed, analisis selanjutnya dapat dilakukan dengan lebih terstruktur dan efisien.

## 3. Modeling

```
# Regresi: Metode ini digunakan untuk memprediksi nilai numerik berdasarkan variabel independen
# Klasifikasi: Ini adalah teknik untuk mengelompokkan data ke dalam kategori yang telah ditentukan.
# Clustering: Teknik ini digunakan untuk mengelompokkan data ke dalam kelompok berdasarkan kesamaan
```

### 3.1 Regression

```
# 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
  )
}
```

**Penjelasan:** Membuat fungsi yang komprehensif untuk membangun dan mengevaluasi model regresi menggunakan Random Forest. Dengan membagi data, melatih model, melakukan prediksi, dan mengevaluasi kinerja, fungsi ini membentuk bagian penting dari proses analisis regresi dalam konteks kinerja akademik siswa. Hasil yang diperoleh akan memberikan wawasan yang berguna untuk memahami faktor-faktor yang mempengaruhi nilai akademis.

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

**Penjelasan:** Model dibangun dan dievaluasi menggunakan data yang telah dipersiapkan sebelumnya (student_perform_regression). Dengan menyimpan hasil dalam student_regression_results.

```
# 3. Display results
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
```

**Penjelasan:** Menampilkan nilai pentingnya fitur, pengguna dapat mengevaluasi dan menginterpretasikan hasil model, serta mengidentifikasi variabel mana yang perlu diperhatikan lebih lanjut dalam konteks kinerja siswa (Dengan nilai: G1=3760.6056; G2=6084.5543; studytime=316.5032; failures=700.0646; freetime=282.1307; goout=333.2419).

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

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

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

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

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

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

**Penjelasan:** Menampilkan Mean Absolute Error (MAE nilai=1.1100683318354) dan Root Mean Squared Error (RMSE nilai=1.79924415122392).

### 3.2 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
  )
}
```

**Penjelasan:** Menyajikan fungsi yang komprehensif untuk membangun dan mengevaluasi model klasifikasi menggunakan pohon keputusan. Dengan membagi data, menangani nilai hilang, melatih model, melakukan prediksi, dan mengevaluasi kinerja, fungsi ini menjadi bagian penting dalam analisis klasifikasi untuk memahami kinerja akademik siswa dalam kategori tertentu (misalnya, kinerja rendah, sedang, atau tinggi).

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

**Penjelasan:** Model dibangun dan dievaluasi menggunakan data yang telah dipersiapkan sebelumnya (student_perform_classification). Dengan menyimpan hasil dalam student_classification_results.

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

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

```
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
```

**Penjelasan:** Menyediakan hasil evaluasi yang komprehensif untuk model klasifikasi yang dibangun. Dengan menampilkan confusion matrix dan statistik terkait, pengguna dapat dengan jelas memahami kinerja model dalam mengklasifikasikan siswa ke dalam kategori kinerja yang berbeda (Hasil=Misalnya, angka 64 di posisi (low, low) menunjukkan bahwa 64 siswa yang sebenarnya berada dalam kategori low berhasil diprediksi dengan benar sebagai low).

### 3.3 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
```

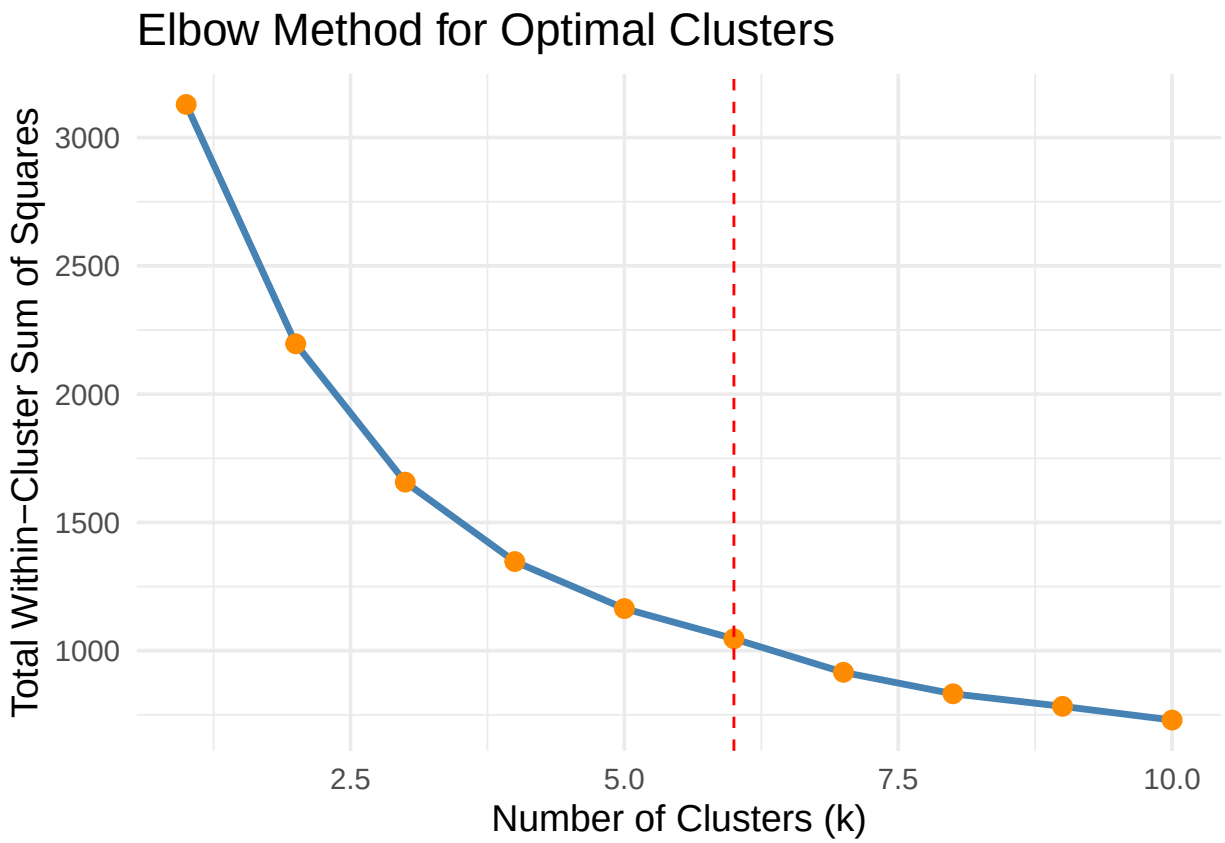
```
  ))
}

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
  ))
}
```

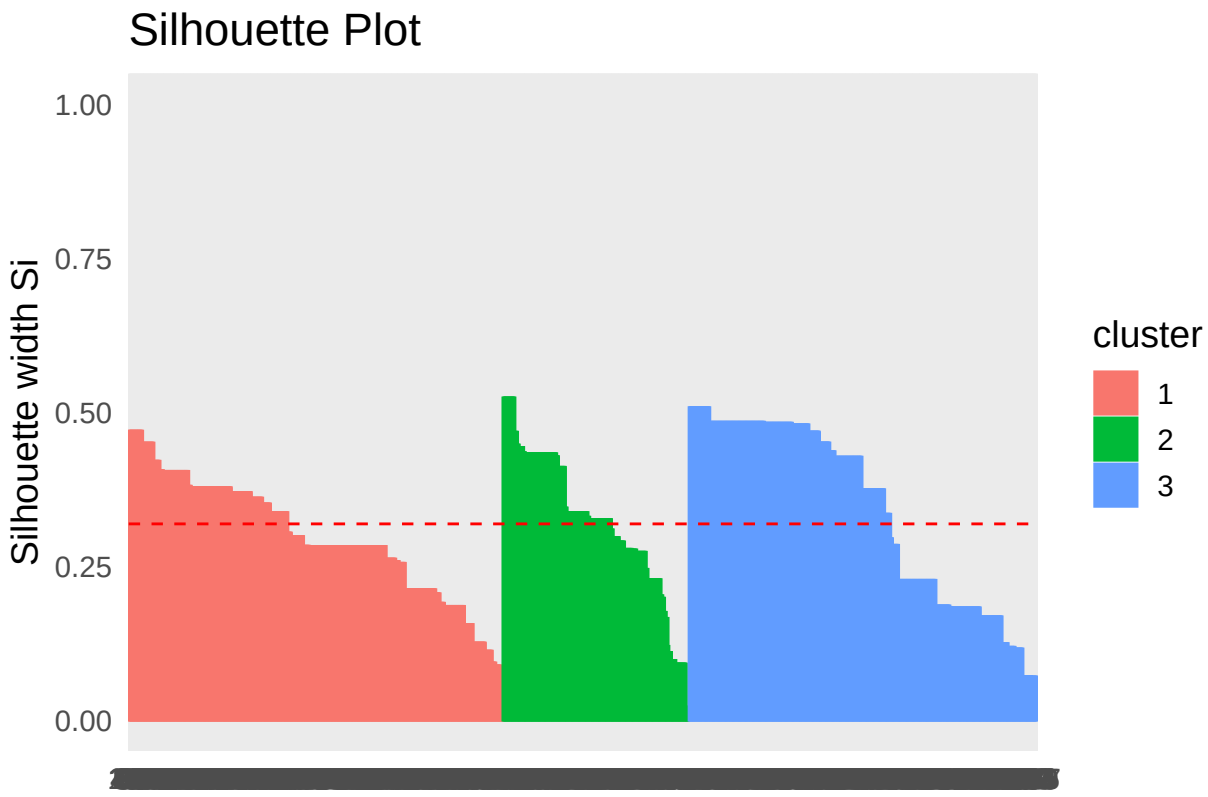
**Penjelasan:** Menyajikan fungsi untuk memberikan analisis kluster yang menyeluruh, mulai dari pemilihan fitur dan normalisasi, hingga visualisasi hasil dan analisis karakteristik kluster. Metode yang ada dalam fungsi: Metode Elbow, K-Means Klustering, Visualisasi PCA (Principal Component Analysis), Matriks Korelasi, Distribusi Kategori Kinerja, Karakteristik Kluster.

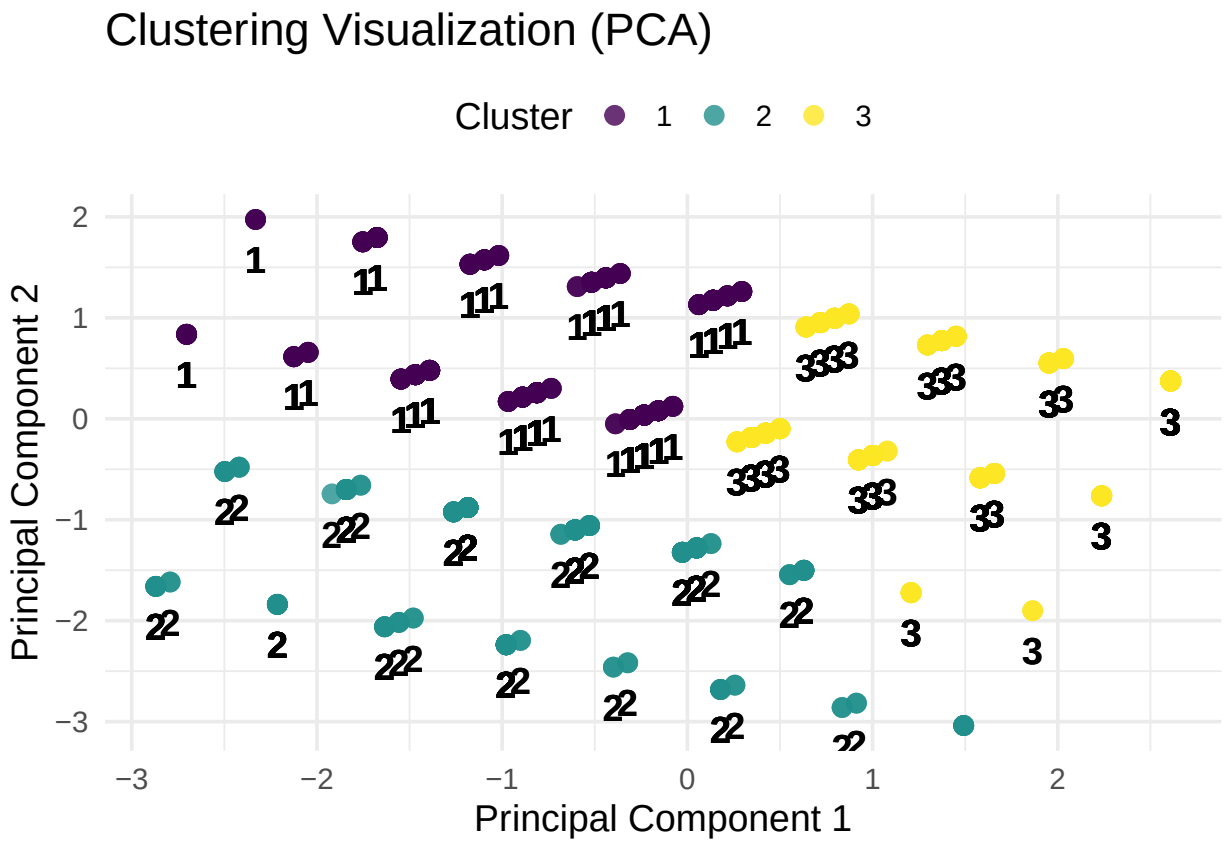
```
# Run clustering analysis
student_clustering_results <- student_perform_clustering_analysis(student_data)
```

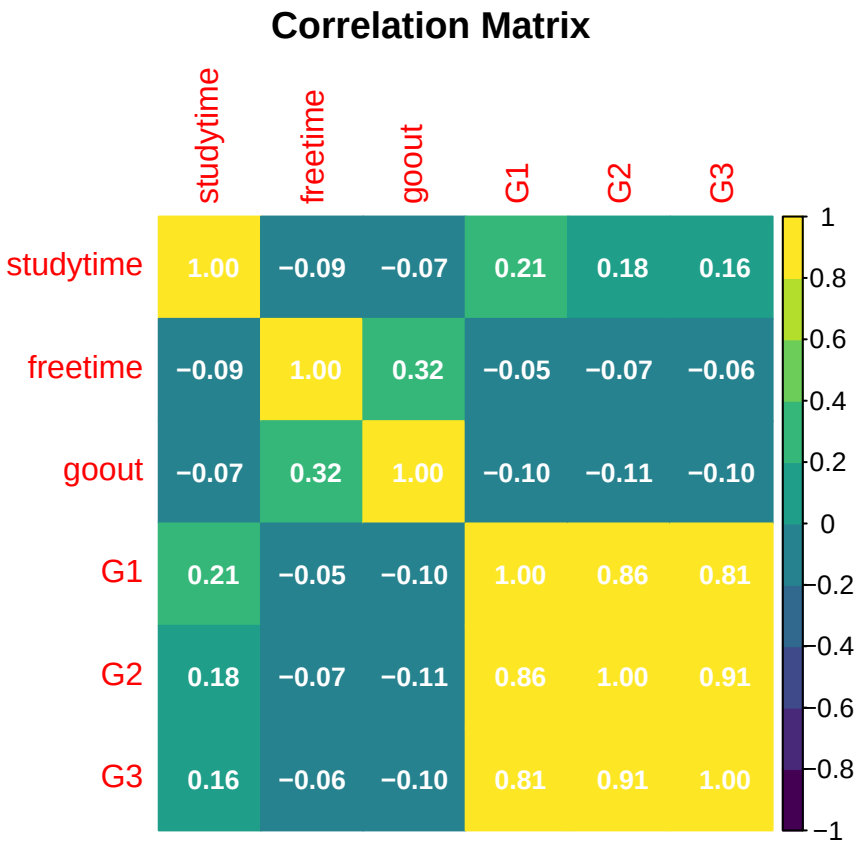
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



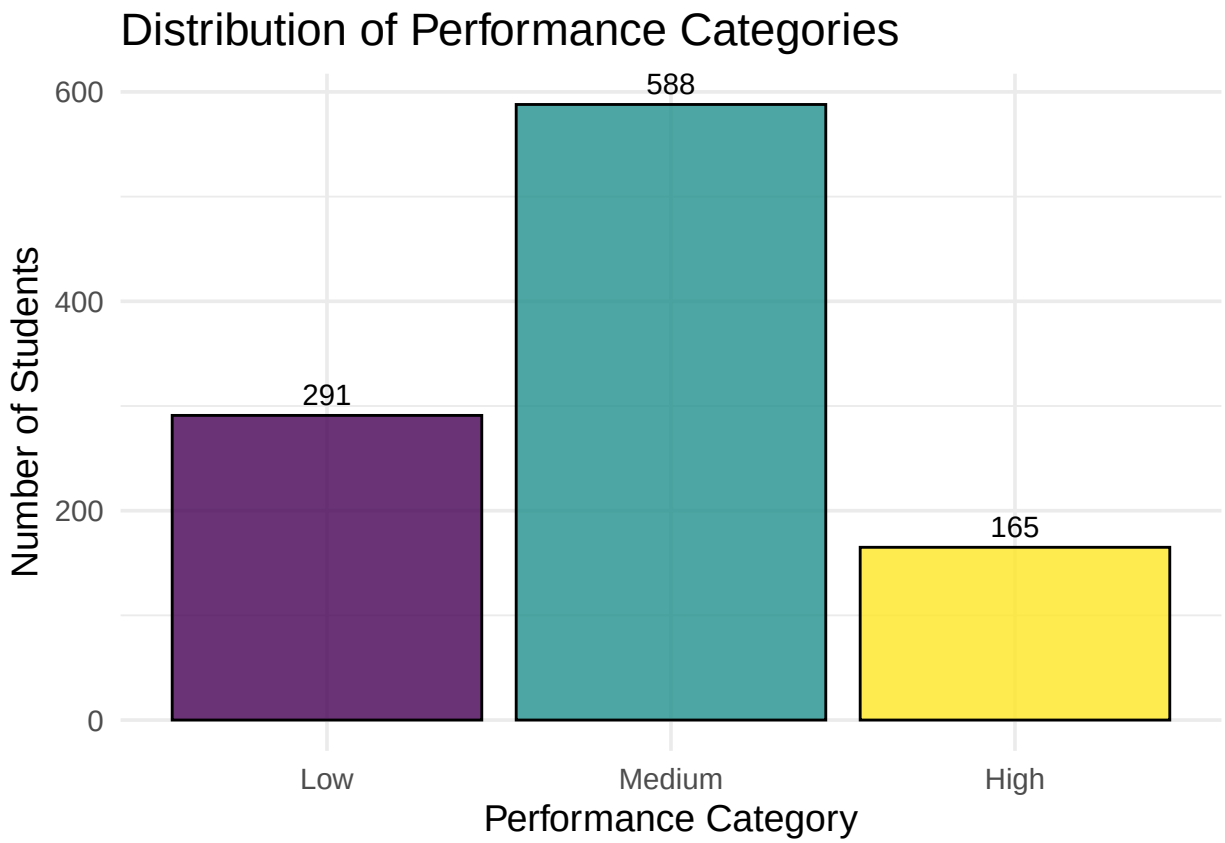

```
##   cluster size ave.sil.width
## 1       1  430          0.30
## 2       2  214          0.32
## 3       3  400          0.34
```


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
```

**Penjelasan:** Berikut penjelasan metode yang digunakan pada clustering:

1. Metode Elbow: Fungsi student_elbow_method digunakan untuk menentukan jumlah optimal kluster (k) dengan menghitung total within-cluster sum of squares (WSS) untuk k dari 1 hingga 10.Dengan visualisasi yang ditingkatkan menggunakan ggplot2, plot ini membantu dalam mengidentifikasi titik "elbow" yang menunjukkan jumlah kluster yang ideal. (Dengan titik "elbow" berada di k=6).

2. K-Means Klustering: Fungsi student_perform_kmeans melakukan klustering K-Means dengan jumlah kluster yang ditentukan (default k = 3). Hasil klustering juga dianalisis menggunakan silhouette plot untuk mengevaluasi seberapa baik klustering dilakukan. (Dengan hasil yang lumayan bagus, setiap cluster positif yang mendekati 1).

3. Visualisasi PCA: Fungsi student_pca_visualization menghasilkan visualisasi dari hasil PCA (Principal Component Analysis) untuk menunjukkan bagaimana data terdistribusi dalam ruang dua dimensi berdasarkan kluster yang dihasilkan. Ini membantu dalam memahami struktur data dan karakteristik setiap kluster. (Dengan PCA yang berada di antara -3 dan 2).

4. Matriks Korelasi: Matriks korelasi dihitung untuk mengeksplorasi hubungan antara variabel studytime, freetime, goout, dan nilai akademik G1, G2, G3. Visualisasi matriks korelasi menggunakan corrplot

untuk memberikan gambaran visual tentang hubungan antar variabel. (Dengan hasil berada diantara -1 dan 1, jika mendekati 1 maka hubungan dari varibel tersebut kuat begitupun sebaliknya).

5. Distribusi Kategori Kinerja: Kategori kinerja siswa dibagi menjadi tiga kelompok: low, medium, dan high berdasarkan nilai G2. Visualisasi distribusi ini menggunakan bar plot untuk menunjukkan jumlah siswa dalam setiap kategori. (Dengan hasil low=291, medium=588, high=165 ).

6. Karakteristik Kluster: Analisis karakteristik kluster dilakukan dengan menghitung rata-rata study-time, freetime, dan goout untuk setiap kluster. Ini memberikan wawasan tentang perilaku dan kebiasaan siswa dalam setiap kluster. (Dengan hasil cluster 1 = mean_studytime(1.630233), mean_freetime(2.634884), mean_goout(2.413953); cluster 2 = mean_studytime(3.289720), mean_freetime(3.009346), mean_goout(2.850467); cluster 3 = mean_studytime(1.630000), mean_freetime(3.912500), mean_goout(4.117500))