

Mini Project Academic Performance Analysis

Deni Permana - Diaz Fahreza Akbar

2024-11-24

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' menampilkan beberapa plot; dan 'reshape2' untuk mengubah format data

Import Dataset

Membaca dataset dari file CSV ke dalam R.

```
# 1. Dataset 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 - 5th to 9th grade, 3 - secondary education or 4 - 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')

No	Nama Variabel	Type	Deskripsi
12	guardian	Categorical	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
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

```
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'

```
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.

```
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.

```
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.


```
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.

```
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
##
##
##      travelttime      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:

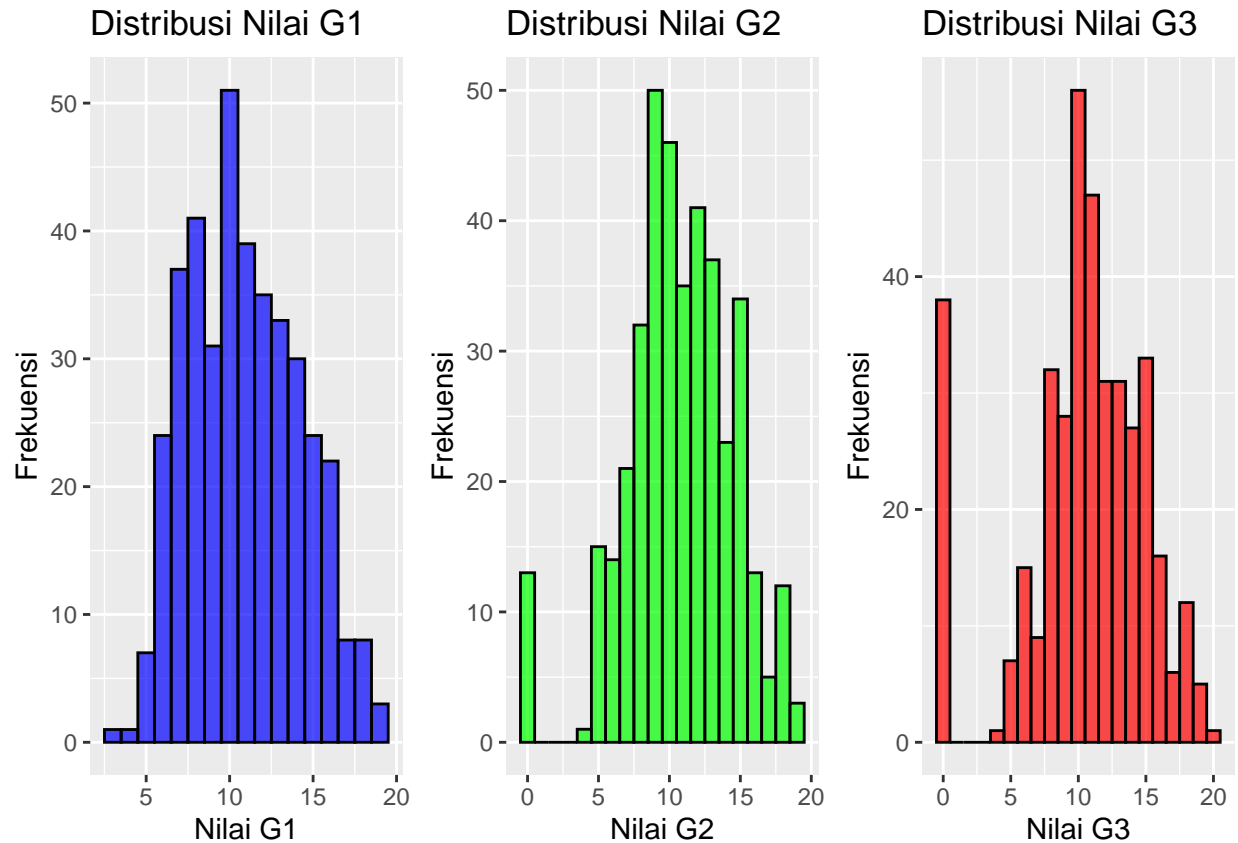
2.3.2.1 Histogram: Histogram untuk Nilai G1, G2, G3

```
# 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")

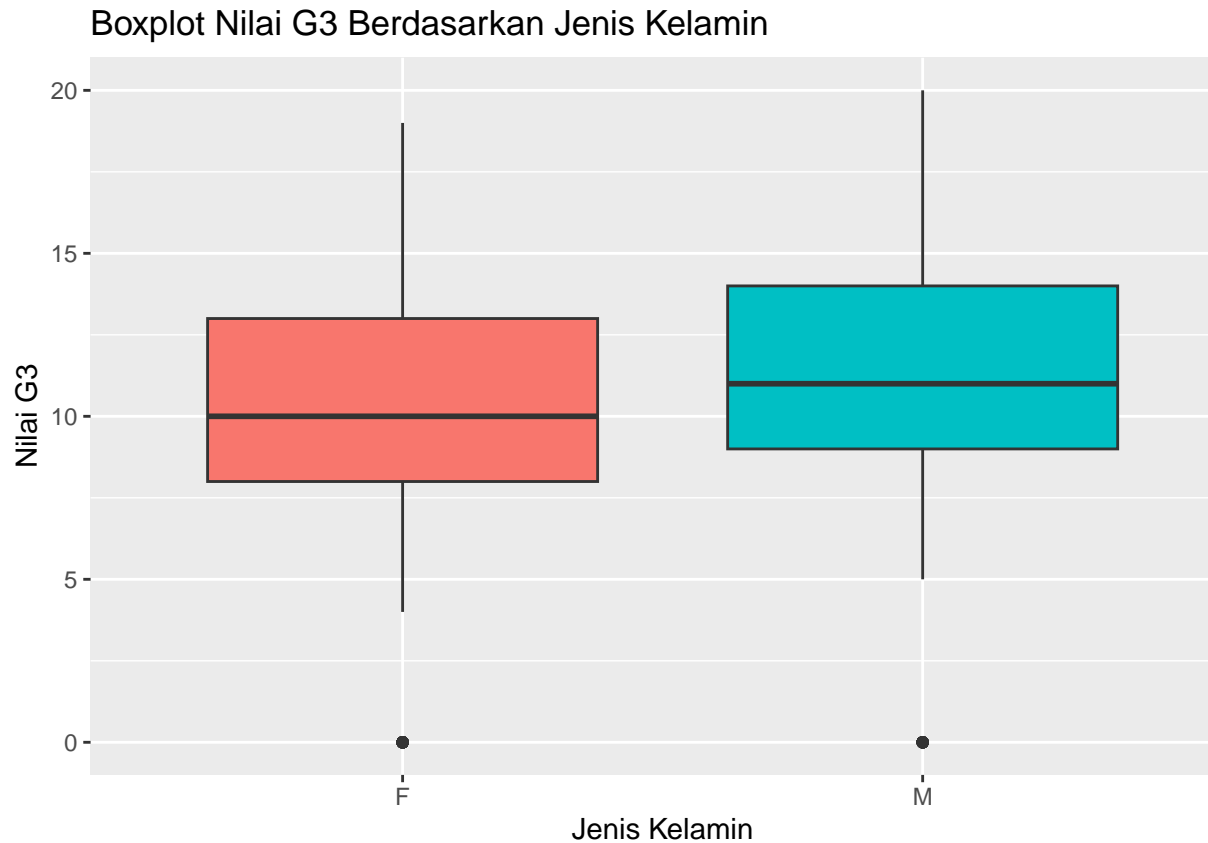
# Tampilkan histogram dalam satu tampilan
grid.arrange(g1_hist, g2_hist, g3_hist, ncol = 3)
```



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)
```

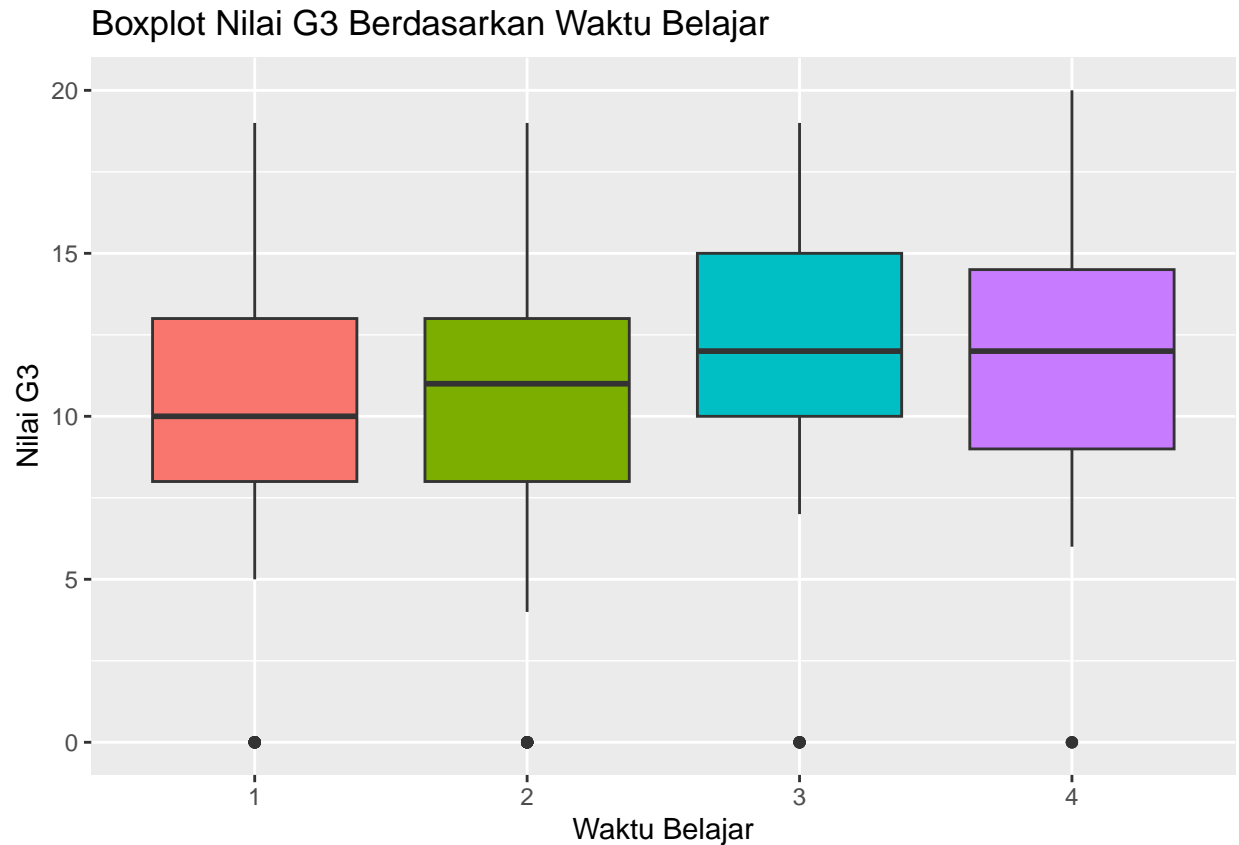


2.3.2.2 Boxplot:

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")

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.

```
# 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 nilainya 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:

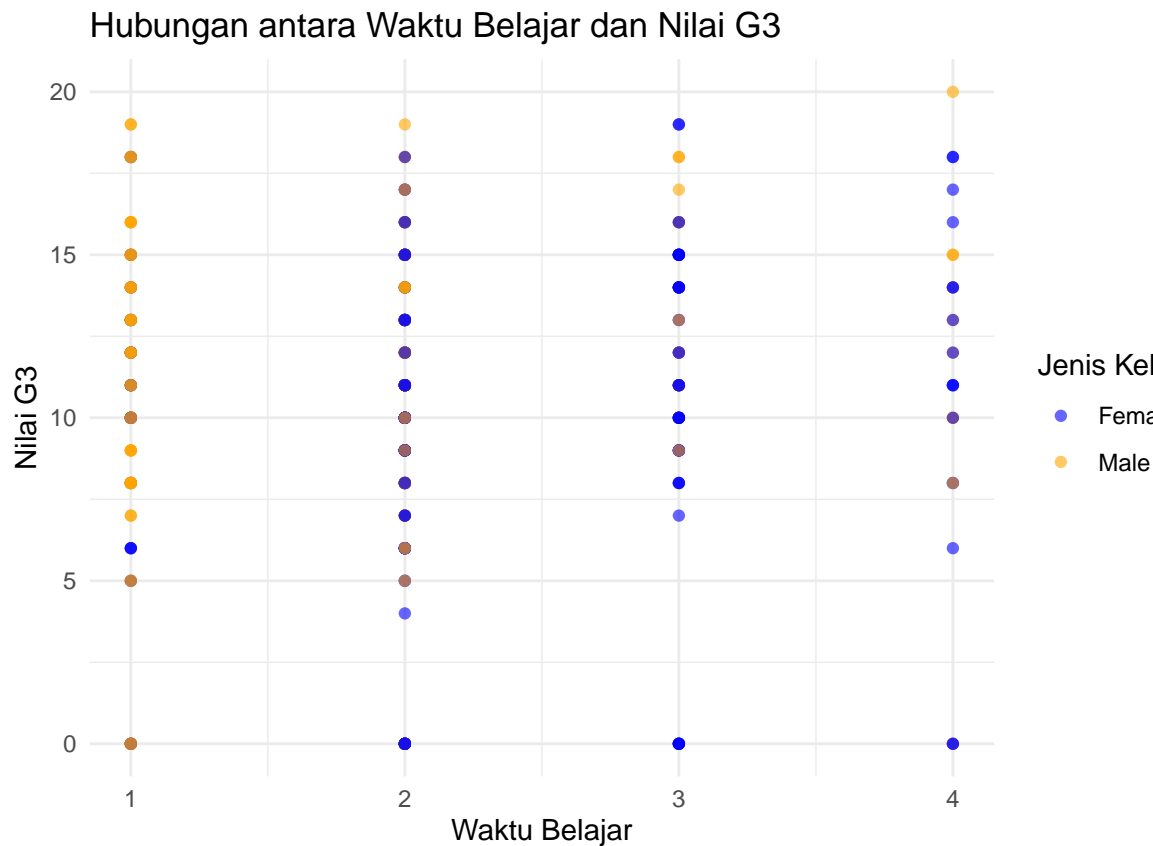
Visualisasi yang jelas dan informatif yang menunjukkan hubungan, tren, dan pola dalam data, serta mendukung narasi analisis.

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)
```



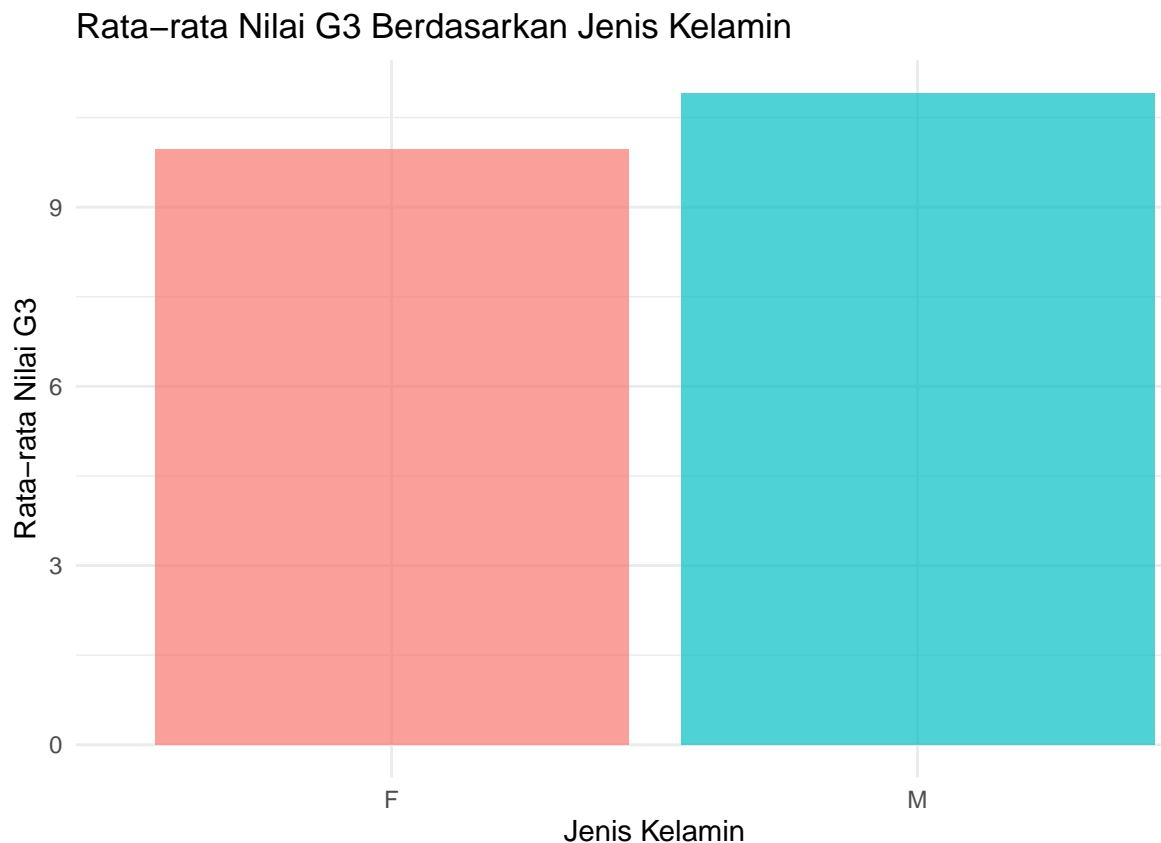
3.3.1.1 Scatter Plot:

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.

```
# 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:

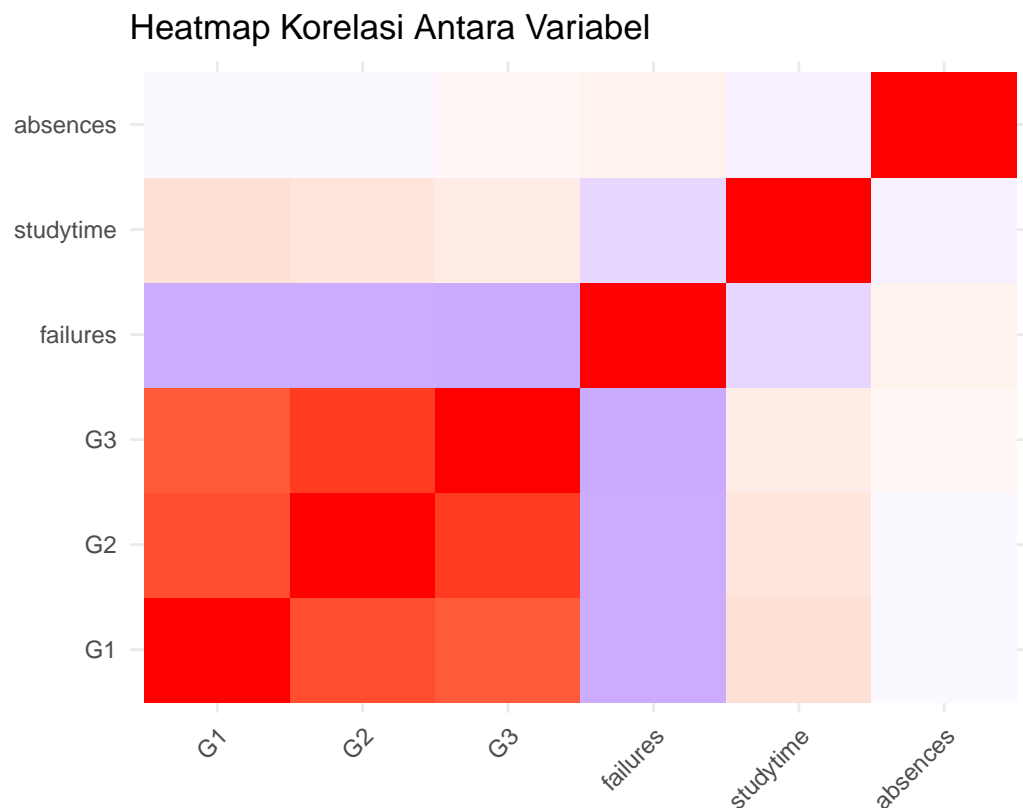
Penjelasan: Bar plot ini menunjukkan bahwa rata-rata nilai G3 untuk siswa laki-laki dan perempuan hampir sama, mengindikasikan bahwa jenis kelamin tidak memiliki pengaruh signifikan terhadap nilai akhir.

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

```
## 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.

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

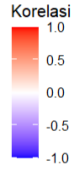
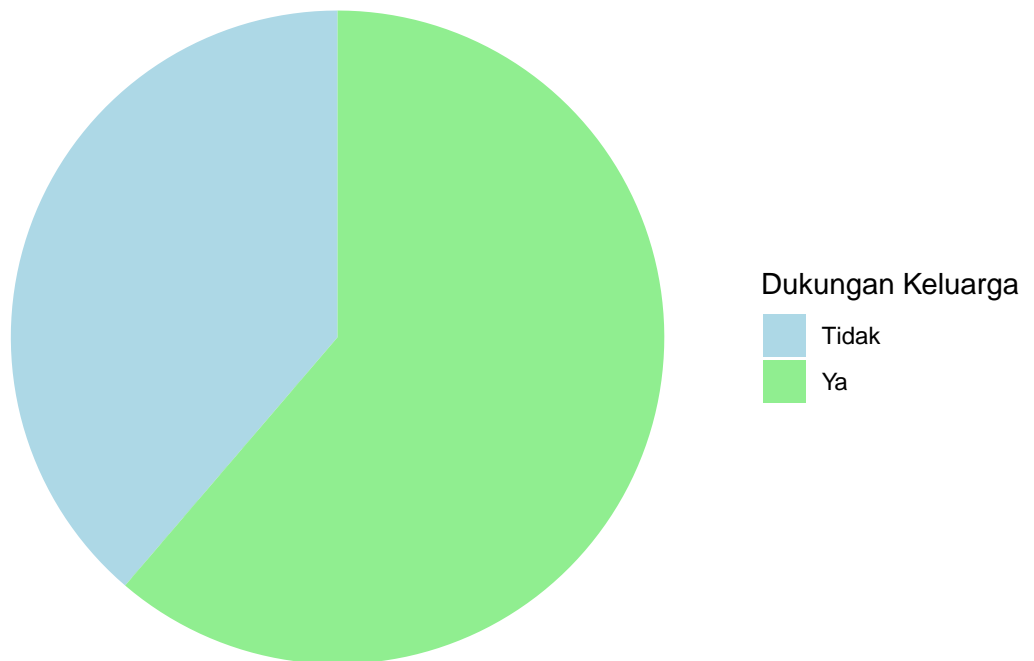


Figure 1: nilai 1.0 menunjukkan korelasi positif yang kuat, nilai -1.0 menunjukkan korelasi negatif yang kuat, dan jika nilai mendekati nol artinya tidak ada korelasi yang jelas. jadi semakin nilainya mendekati nilai positif maka tren nya positif begitupun sebaliknya.

```
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)
```

Proporsi Dukungan Keluarga



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.