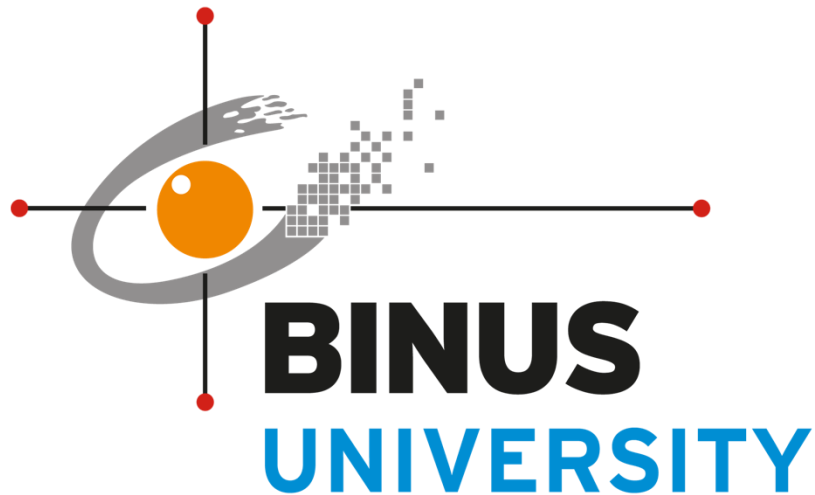


Analisa Data
“Student Grade Prediction”



Disusun oleh:
Patrick Jonathan
2440064791

JURUSAN TEKNIK INFORMATIKA DAN STATISTIKA
PROGRAM STUDI DATA MINING AND VISUALIZATION
UNIVERSITAS BINA NUSANTARA

2022

INTRODUCTION

Pendahuluan

Dataset yang saya pakai untuk dianalisis diambil dari kaggle.com yang berjudul Student Grade Prediction (<https://www.kaggle.com/dipam7/student-grade-prediction>). Dataset ini memiliki 2 tipe data, yaitu data numerik dan data kategorikal. Dataset ini berisi berbagai variabel yang cukup bervariasi seperti jenis kelamin, umur, alamat, jumlah anggota keluarga, waktu belajar, waktu berpergian, pekerjaan orang tua, dan masih banyak lagi. Berdasarkan dataset tersebut, saya akan menganalisa hubungan antara beberapa variabel yang memiliki hubungan dengan nilai akhir siswa tersebut serta memprediksi nilai akhir siswa tersebut.

Tujuan

1. Untuk memenuhi Ujian Tengah Semester Mata Kuliah Data Mining and Visualization
2. Melakukan analisis dan memvisualisasikan hasil analisis dengan menggunakan program Rstudio
3. Menganalisa hubungan antara beberapa variabel yang memiliki hubungan dengan nilai akhir siswa tersebut serta memprediksi nilai akhir siswa tersebut
4. Menampilkan model prediksi regresi yang dibuat

DATA DESCRIPTION

Dataset ini memiliki 395 data dengan 33 variabel yang berupa informasi yang relevan tentang siswa, seperti jenis kelamin, umur, alamat, jumlah anggota keluarga, waktu belajar, waktu berpergian, pekerjaan orang tua, dan masih banyak lagi.

Berikut adalah beberapa variabel berupa informasi yang relevan tentang siswa :

1. school – nama sekolah siswa ('GP' : Gabriel Pereira atau 'MS' : Mousinho da Silveira)
2. sex – jenis kelamin siswa ('F' : perempuan atau 'M' – laki-laki)
3. age – umur siswa
4. address – tipe tempat tinggal siswa ('U' - urban atau 'R' - rural)
5. famsize – ukuran keluarga siswa ('LE3' : lebih sedikit dari sama dengan 3 atau 'GT3' : lebih banyak dari 3)
6. Pstatus – status tinggal bersama keluarga ('T' : tinggal bersama atau 'A' : tidak tinggal bersama)
7. Medu – edukasi ibu (0 : tidak ada, 1 : pendidikan dasar (kelas 4), 2 : kelas 5 – kelas 9, 3 : kelas 9 – kelas 12 atau 4 : kuliah)
8. Fedu – edukasi ayah (0 : tidak ada, 1 : pendidikan dasar (kelas 4), 2 : kelas 5 – kelas 9, 3 : kelas 9 – kelas 12 atau 4 : kuliah)
9. Mjob – pekerjaan ibu
10. Fjob – pekerjaan ayah
11. reason – alasan memilih sekolah tersebut
12. guardian – wali siswa
13. traveltime – perkiraan waktu dari rumah ke sekolah (1 : <15 menit, 2 : 15 sampai 30 menit, 3 : 30 menit sampai 1 jam, atau 4 : >1 jam)
14. studytime – jumlah waktu yang dihabiskan untuk belajar dalam seminggu (1 : <15 menit, 2 : 15 sampai 30 menit, 3 : 30 menit sampai 1 jam, atau 4 : >1 jam)
15. failures – jumlah mata pelajaran yang gagal (mulai dari 1 sampai dengan 4)
16. schoolsup – mengidentifikasi apakah siswa memiliki dukungan pendidikan extra (yes atau no)
17. famsup - mengidentifikasi apakah siswa memiliki dukungan pendidikan di keluarga (yes atau no)
18. paid – mengidentifikasi apakah siswa mengikuti kelas berbayar extra (yes atau no)

19. activities – mengidentifikasi apakah siswa mengikuti kegiatan ekstrakurikuler (yes atau no)
20. nursery – mengidentifikasi apakah siswa pernah bersekolah di taman kanak-kanak (yes atau no)
21. higher – mengidentifikasi apakah siswa ingin melanjutkan ke jenjang yang lebih tinggi (yes atau no)
22. internet – mengidentifikasi apakah siswa memiliki akses internet di rumah (yes atau no)
23. romantic – mengidentifikasi apakah siswa memiliki hubungan romantis (yes atau no)
24. famrel – kualitas hubungan keluarga (mulai dari 1 : sangat buruk sampai dengan 5 : sangat baik)
25. freetime - waktu luang sepulang sekolah (mulai dari 1 : tidak ada sampai dengan 5 : sangat banyak)
26. goout – frekuensi berpergian bersama teman (mulai dari 1 : tidak pernah sampai dengan 5 : sangat sering)
27. Dalc – frekuensi konsumsi alkohol pada hari biasa (mulai dari 1 : tidak pernah sampai dengan 5 : sangat sering)
28. Walc – frekuensi konsumsi alkohol pada akhir pekan (mulai dari 1 : very low sampai dengan 5 : very high)
29. health – status kesehatan saat ini (mulai dari 1 : sangat buruk sampai dengan 5 : sangat baik)
30. absences – jumlah absen pada saat sekolah (mulai dari 0 sampai dengan 93)
31. G1 – nilai ujian pertama (mulai dari 0 sampai dengan 20)
32. G2 – nilai ujian kedua (mulai dari 0 sampai dengan 20)
33. G3 – nilai akhir (mulai dari 0 sampai dengan 20)

DATA EXPLORATION AND VISUALIZATION

Exploratory Data Analysis

Pertama-tama, kita perlu mengimport library yang akan digunakan terlebih dahulu

```
> # Import library yang akan digunakan
> library(dplyr)
```

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(skimr)
> library(ggplot2)
> library(ggpubr)
> library(corrplot)
corrplot 0.92 loaded
```

Selanjutnya, kita import dataset yang akan digunakan yaitu student-mat

```
> setwd("C:\\Users\\Patrick Jonathan\\Documents\\Patrick Jonathan\\Sem 3\\UAS\\Data Mining")
> studentDataSet <- read.csv("student-mat.csv")
```

Dataset ini saya masukkan ke dalam suatu data frame dengan nama studentDataSet

Selanjutnya kita dapat melihat banyaknya keseluruhan data dan total variabelnya dengan menggunakan fungsi **dim()** sehingga kita dapat melihat dimensi dari data frame tersebut

```
> # Melihat banyaknya data keseluruhan dan total variabel nya
> dim(studentDataSet)
[1] 395 33
```

Dapat dilihat bahwa dataset ini memiliki total 395 data dengan 33 variabel yang berbeda

Selanjutnya, kita dapat melihat atau menampilkan 5 data teratas dan juga 5 data terbawah dengan menggunakan fungsi **head()** dan **tail()**

















```
> # Menampilkan 5 data teratas dan 5 data terbawah
> head(studentDataSet,5)
  school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup paid
1 GP F 18 U GT3 A 4 4 at_home teacher course mother 1 2 0 yes no no
2 GP F 17 U GT3 T 1 1 at_home other course father 1 2 0 no yes no
3 GP F 15 U LE3 T 1 1 at_home other other mother 1 2 3 yes no yes
4 GP F 15 U GT3 T 4 2 health services home mother 1 3 0 no yes yes
5 GP F 16 U GT3 T 3 3 other other home father 1 2 0 no yes yes
  activities nursery higher internet romantic famrel freetime goout dalc walc health absences G1 G2 G3
1 no yes yes no no 4 3 4 1 1 3 6 5 6 6
2 no no yes yes no 5 3 3 1 1 3 4 5 5 6
3 no yes yes yes no 4 3 2 2 3 3 10 7 8 10
4 yes yes yes yes yes 3 2 2 1 1 5 2 15 14 15
5 no yes yes no no 4 3 2 1 2 5 4 6 10 10
> tail(studentDataSet,5)
  school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason guardian traveltime studytime failures schoolsup famsup paid
391 MS M 20 U LE3 A 2 2 services services course other 1 2 2 no yes yes
392 MS M 17 U LE3 T 3 1 services services course mother 2 1 0 no no no
393 MS M 21 R GT3 T 1 1 other other course mother 1 1 3 no no no
394 MS M 18 R LE3 T 3 2 services other course mother 3 1 0 no no no
395 MS M 19 U LE3 T 1 1 other at_home course father 1 1 0 no no no
  activities nursery higher internet romantic famrel freetime goout dalc walc health absences G1 G2 G3
391 no yes yes no no 5 5 4 4 5 4 11 9 9 9
392 no no yes yes no 2 4 5 3 4 2 3 14 16 16
393 no no yes no no 5 5 3 3 3 3 3 10 8 7
394 no no yes yes no 4 4 1 3 4 5 0 11 12 10
395 no yes yes yes no 3 2 3 3 3 5 5 8 9 9
> |
```

Kemudian saya ingin melihat keseluruhan variabel data

```

> skim(studentDataSet)
-- Data Summary -----
Name                               values
Number of rows                    395
Number of columns                  33
Column type frequency:
  character                        17
  numeric                         16
Group variables                    None

-- Variable type: character -----
# A tibble: 17 x 8
  skim_variable n_missing complete_rate min max empty n_unique whitespace
* <chr>         <int>         <dbl> <int> <int> <int>   <int>   <int>
1 school         0             1     2     2     0     2     0
2 sex            0             1     1     1     0     2     0
3 address        0             1     1     1     0     2     0
4 famsize        0             1     3     3     0     2     0
5 Pstatus        0             1     1     1     0     2     0
6 Mjob           0             1     5     8     0     5     0
7 Fjob           0             1     5     8     0     5     0
8 reason         0             1     4    10     0     4     0
9 guardian       0             1     5     6     0     3     0
10 schoolsup      0             1     2     3     0     2     0
11 famsup         0             1     2     3     0     2     0
12 paid           0             1     2     3     0     2     0
13 activities     0             1     2     3     0     2     0
14 nursery        0             1     2     3     0     2     0
15 higher         0             1     2     3     0     2     0
16 internet       0             1     2     3     0     2     0
17 romantic       0             1     2     3     0     2     0

-- Variable type: numeric -----
# A tibble: 16 x 11
  skim_variable n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist
* <chr>         <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 age           0             1 16.7  1.28 15  16  17  18  22  
2 Medu          0             1  2.75 1.09  0   2   3   4   4  
3 Fedu          0             1  2.52 1.09  0   2   2   3   4  
4 traveltime    0             1  1.45 0.698  1   1   1   2   4  
5 studytime     0             1  2.04 0.839  1   1   2   2   4  
6 failures      0             1  0.334 0.744  0   0   0   0   3  
7 famrel        0             1  3.94 0.897  1   4   4   5   5  
8 freetime      0             1  3.24 0.999  1   3   3   4   5  
9 goout         0             1  3.11 1.11  1   2   3   4   5  
10 Dalc          0             1  1.48 0.891  1   1   1   2   5  
11 Walc          0             1  2.29 1.29  1   1   2   3   5  
12 health        0             1  3.55 1.39  1   3   4   5   5  
13 absences      0             1  5.71 8.00  0   0   4   8  75  
14 G1            0             1 10.9  3.32  3   8  11  13  19  
15 G2            0             1 10.7  3.76  0   9  11  13  19  
16 G3            0             1 10.4  4.58  0   8  11  14  20  

```

Dengan menggunakan fungsi **skim()**, kita dapat melihat keseluruhan variabel yang ada, bahkan sudah dikelompokkan. Variabel kategorikal antara lain ada school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, dan romatic. Sedangkan untuk variabel numerik antara lain ada age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, dan G3. Selain itu kita juga dapat melihat missing value pada setiap variabel. Berdasarkan tabel tersebut seluruh variabel tidak memiliki missing value sama sekali. Kita juga dapat melihat rata-rata, standar deviasi, nilai minimum, median, serta nilai maksimum dari variabel numerik.

- Mengecek missing value (NA)

```

> # Check missing value dalam data
> colSums(is.na(studentDataSet))
  school    sex    age  address  famsize  Pstatus    Medu    Fedu    Mjob    Fjob    reason  guardian
traveltime studytime failures schoolsup  famsup    paid activities  nursery  higher  internet  romantic  famrel
  freetime    goout    dalc    walc    health  absences    G1    G2    G3
0
> |

```

Berdasarkan tabel diatas, dapat dilihat bahwa setiap variabel tidak memiliki missing value, sehingga dataset cukup baik untuk dianalisis

- Mengecek adanya duplicate data

```

> # Check duplicate data
> sum(duplicated(studentDataSet))
[1] 0
> |

```

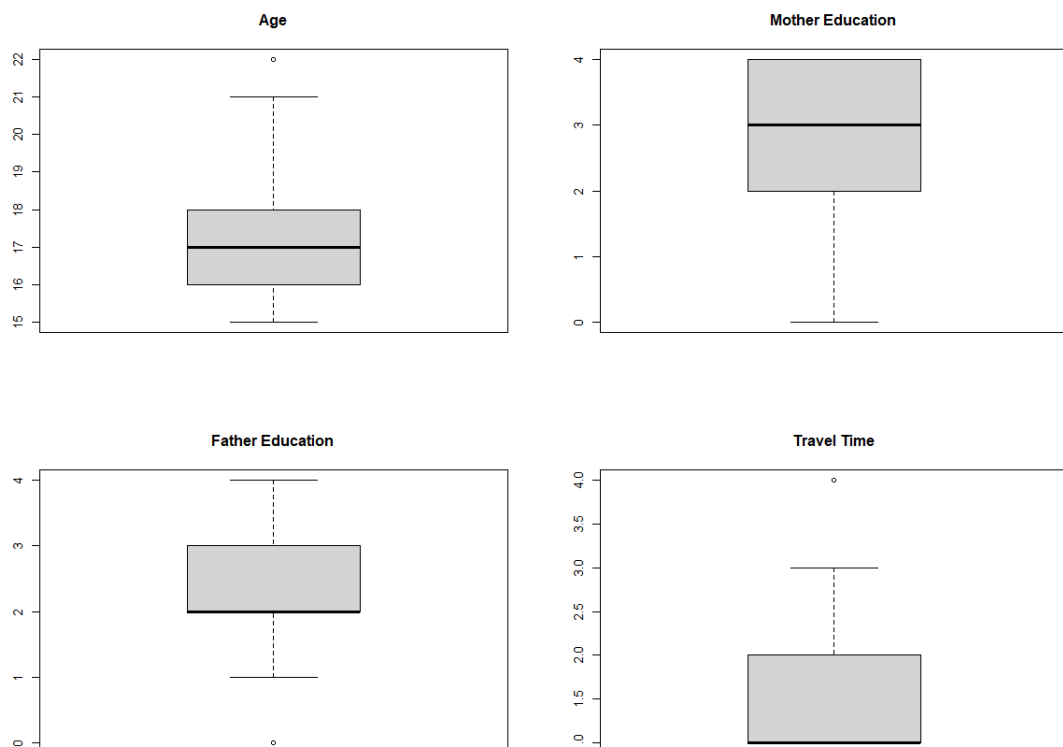
Dataset ini juga tidak memiliki duplicate data

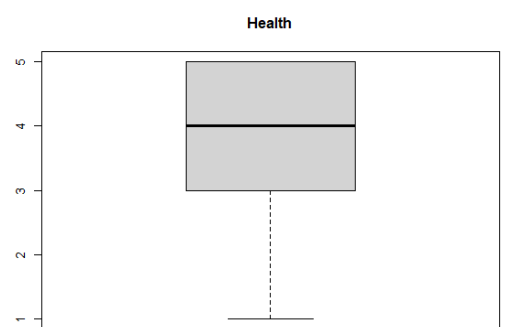
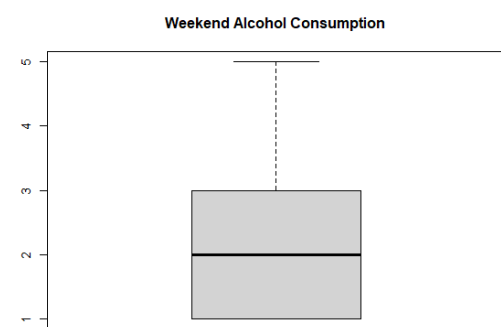
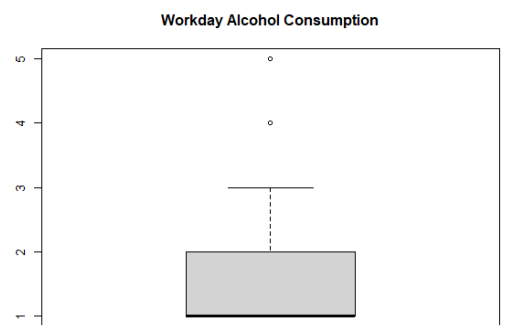
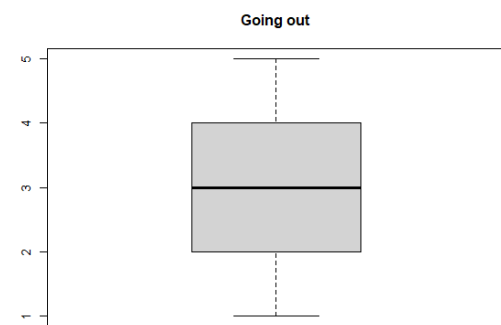
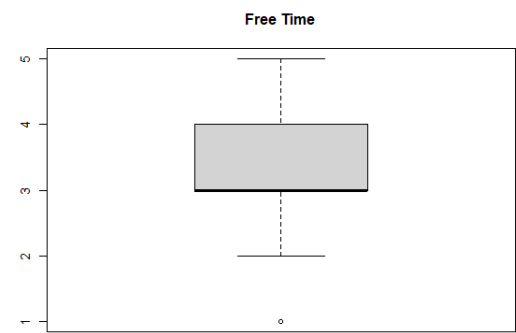
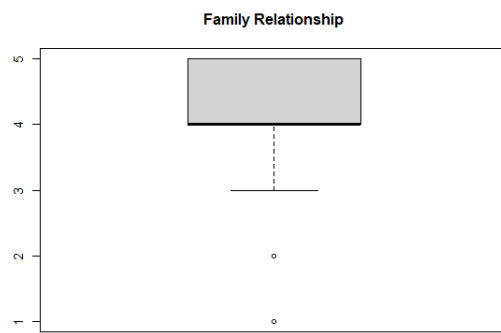
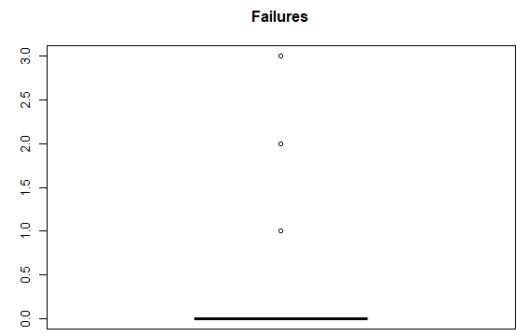
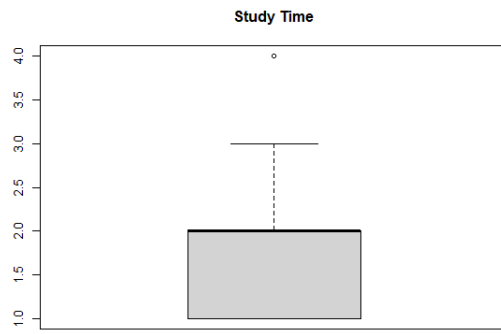
- Mengecek adanya outlier

```

> # Check adanya outlier dalam data numerik
> par(mfrow=c(2,2))
> boxplot(studentDataSet$age, main = "Age")
> boxplot(studentDataSet$Medu, main = "Mother Education")
> boxplot(studentDataSet$Fedu, main = "Father Education")
> boxplot(studentDataSet$traveltime, main = "Travel Time")
> boxplot(studentDataSet$studytime, main = "Study Time")
> boxplot(studentDataSet$failures, main = "Failures")
> boxplot(studentDataSet$famrel, main = "Family Relationship")
> boxplot(studentDataSet$freetime, main = "Free Time")
> boxplot(studentDataSet$goout, main = "Going out")
> boxplot(studentDataSet$dalc, main = "Workday Alcohol Consumption")
> boxplot(studentDataSet$walc, main = "Weekend Alcohol Consumption")
> boxplot(studentDataSet$health, main = "Health")
> boxplot(studentDataSet$absences, main = "Absences")
> boxplot(studentDataSet$G1, main = "Grade 1")
> boxplot(studentDataSet$G2, main = "Grade 2")
> boxplot(studentDataSet$G3, main = "Final Grade")
> |

```

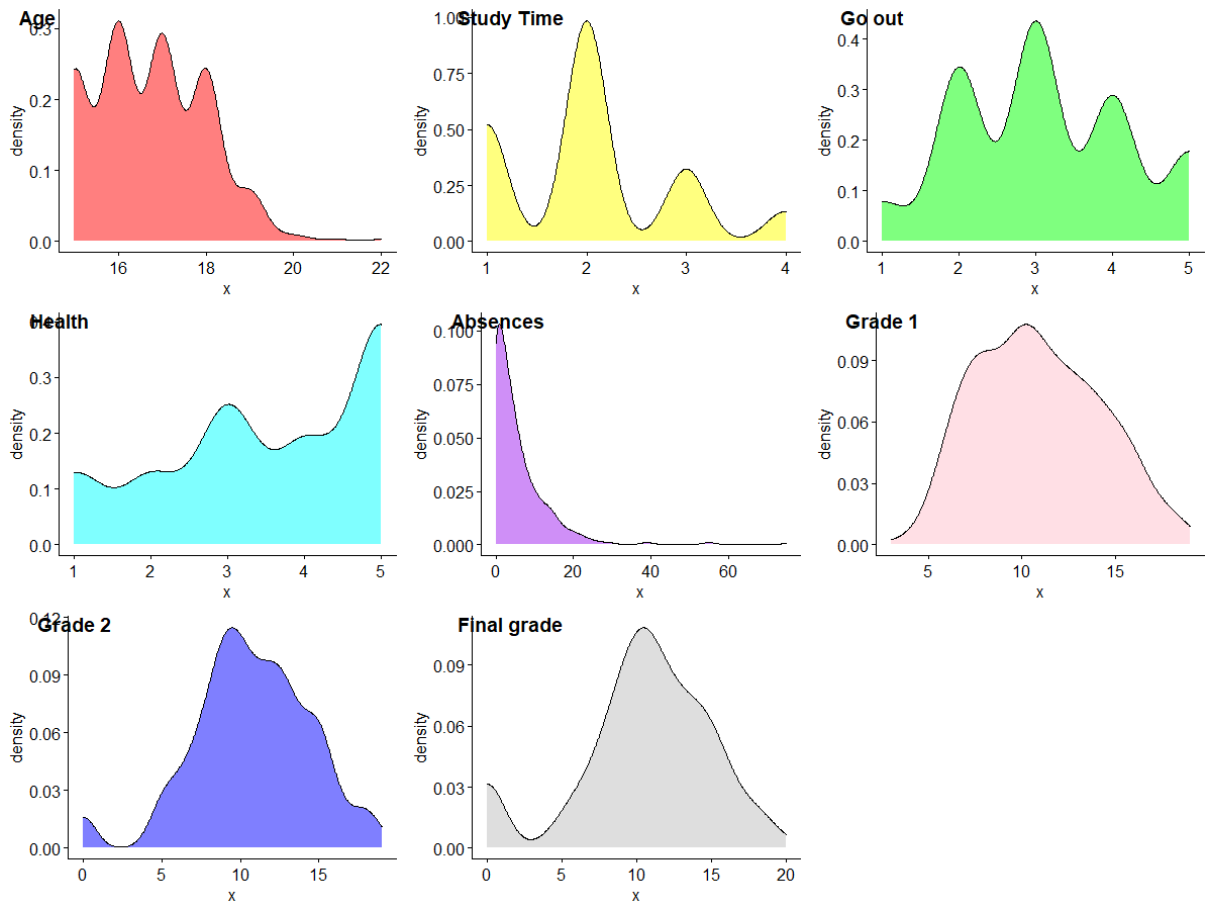





```

> # Menampilkan Density Plot
> ggarrange(ggdensity(studentDataset$age, fill = 'red'),
+           ggdensity(studentDataset$studytime, fill = 'yellow'),
+           ggdensity(studentDataset$goout, fill = 'green'),
+           ggdensity(studentDataset$health, fill = 'cyan'),
+           ggdensity(studentDataset$absences, fill = 'purple'),
+           ggdensity(studentDataset$G1, fill = 'pink'),
+           ggdensity(studentDataset$G2, fill = 'blue'),
+           ggdensity(studentDataset$G3, fill = 'grey'),
+           labels= c("Age", "Study Time", "Go out", "Health", "Absences", "Grade 1", "Grade 2", "Final grade"))
>

```



Pada gambar diatas dapat disimpulkan bahwa variabel studytime, goout, health, G1, G2, dan G3 mendekati distribusi normal karena grafiknya tidak condong ke arah kiri maupun ke kanan (no skew). Sedangkan grafik pada variabel age, dan absences agak condong ke arah kanan (right-skewed) sehingga tidak terdistribusi secara normal.

- **Korelasi antara setiap variabel numerik**

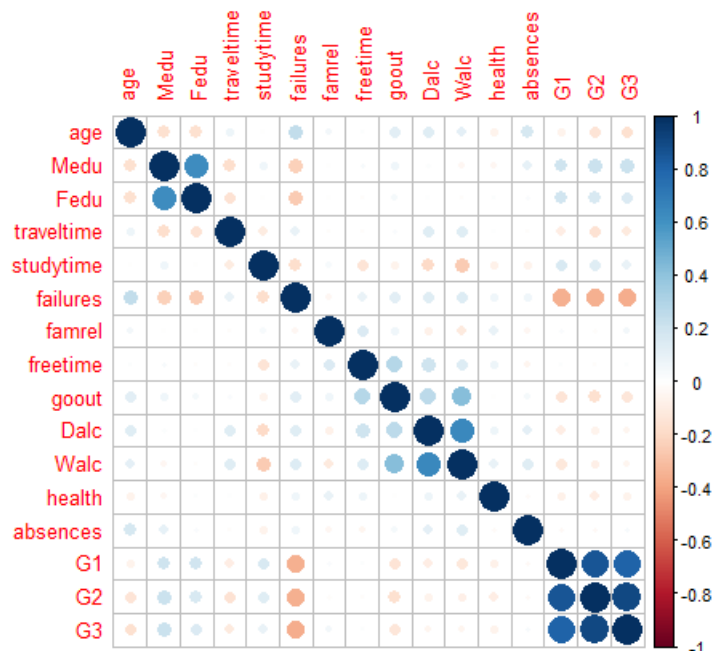
Korelasi antar setiap variabel numerik dapat dilihat dengan menggunakan fungsi **cor()**

```
> cor(studentDataSet[,unlist(lapply(studentDataSet, is.numeric))])
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc
age	1.000000000	-0.163658419	-0.163438069	0.070640721	-0.004140037	0.24366538	0.053940096	0.01643439	0.126963880	0.131124605
Medu	-0.163658419	1.000000000	0.623455112	-0.171639305	0.064944137	-0.23667996	-0.003914458	0.03089087	0.064094438	0.019834099
Fedu	-0.163438069	0.623455112	1.000000000	-0.158194054	-0.009174639	-0.25040844	-0.001369727	-0.01284553	0.043104668	0.002386429
traveltime	0.070640721	-0.171639305	-0.158194054	1.000000000	-0.100909119	0.09223875	-0.016807986	-0.01702494	0.028539674	0.138325309
studytime	-0.004140037	0.064944137	-0.009174639	-0.100909119	1.000000000	-0.17356303	0.039730704	-0.14319841	-0.063903675	-0.196019263
failures	0.24366538	-0.23667996	-0.25040844	0.09223875	-0.17356303	1.000000000	-0.044336626	0.09198747	0.124560922	0.136046931
famrel	0.053940096	-0.003914458	-0.001369727	-0.016807986	0.039730704	-0.044336626	1.000000000	0.15070144	0.064568411	-0.077594357
freetime	0.016434389	0.030890867	-0.012845528	-0.017024944	-0.143198407	0.09198747	0.150701444	1.000000000	0.285018715	0.209000848
goout	0.126963880	0.064094438	0.043104668	0.028539674	-0.063903675	0.12456092	0.064568411	0.28501871	1.000000000	0.266993848
Dalc	0.131124605	0.019834099	0.002386429	0.138325309	-0.196019263	0.13604693	-0.077594357	0.20900085	0.266993848	1.000000000
Walc	0.117276052	-0.047123460	-0.012631018	0.134115752	-0.253784731	0.14196203	-0.113397308	0.14782181	0.420385745	0.647544230
health	-0.062187369	-0.046877829	0.014741537	0.007500606	-0.075615863	0.06582728	0.094055728	0.07573336	-0.009577254	0.077179582
absences	0.175230079	0.100284818	0.024472887	-0.012943775	-0.062700175	0.06372583	-0.044354095	-0.05807792	0.044302220	0.111908026
G1	-0.064081497	0.205340997	0.190269936	-0.093039992	0.160611915	-0.35471761	0.022168316	0.01261293	-0.149103967	-0.094158792
G2	-0.143474049	0.215527168	0.164893393	-0.153197963	0.135879999	-0.35589563	-0.018281347	-0.01377714	-0.162250034	-0.064120183
G3	-0.161579438	0.217147496	0.152456939	-0.117142053	0.097819690	-0.36041494	0.051363429	0.01130724	-0.132791474	-0.054660041

```
> |
```

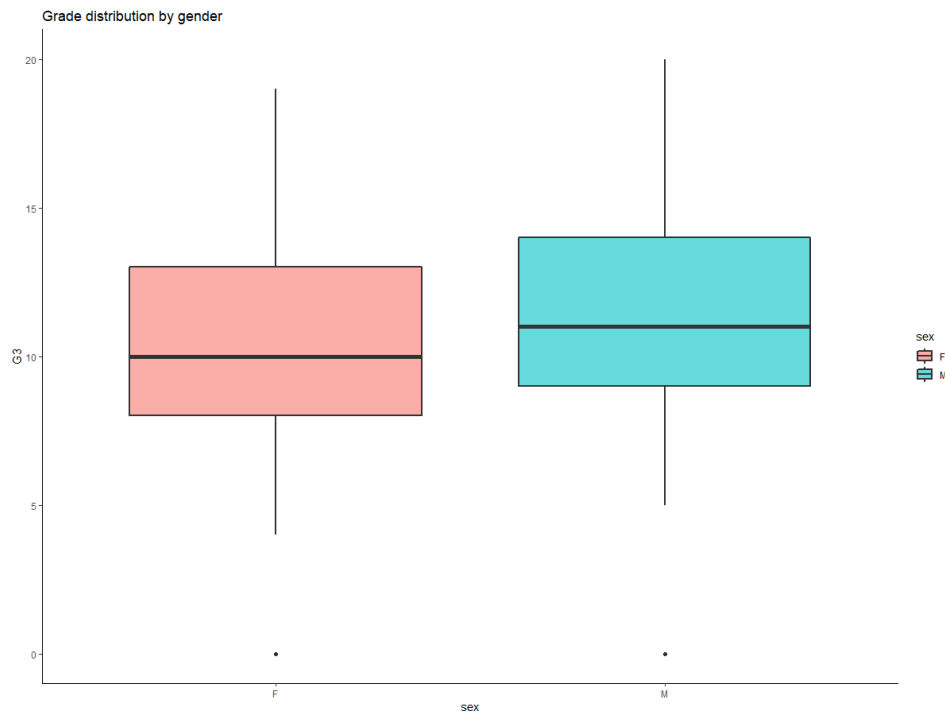
```
> corrplot(cor(studentDataSet[,unlist(lapply(studentDataSet, is.numeric))]))
```



Berdasarkan kedua hasil diatas, dapat disimpulkan bahwa hampir seluruh korelasi antar variabel numeriknya menunjukkan angka yang cukup rendah. Namun korelasi antara G1 dan G2, G1 dan G3, serta G2 dan G3 menunjukkan angka yang cukup besar. Ini artinya hubungan antara ketiga variabel tersebut sangat kuat dan berbanding lurus terhadap satu sama lain.

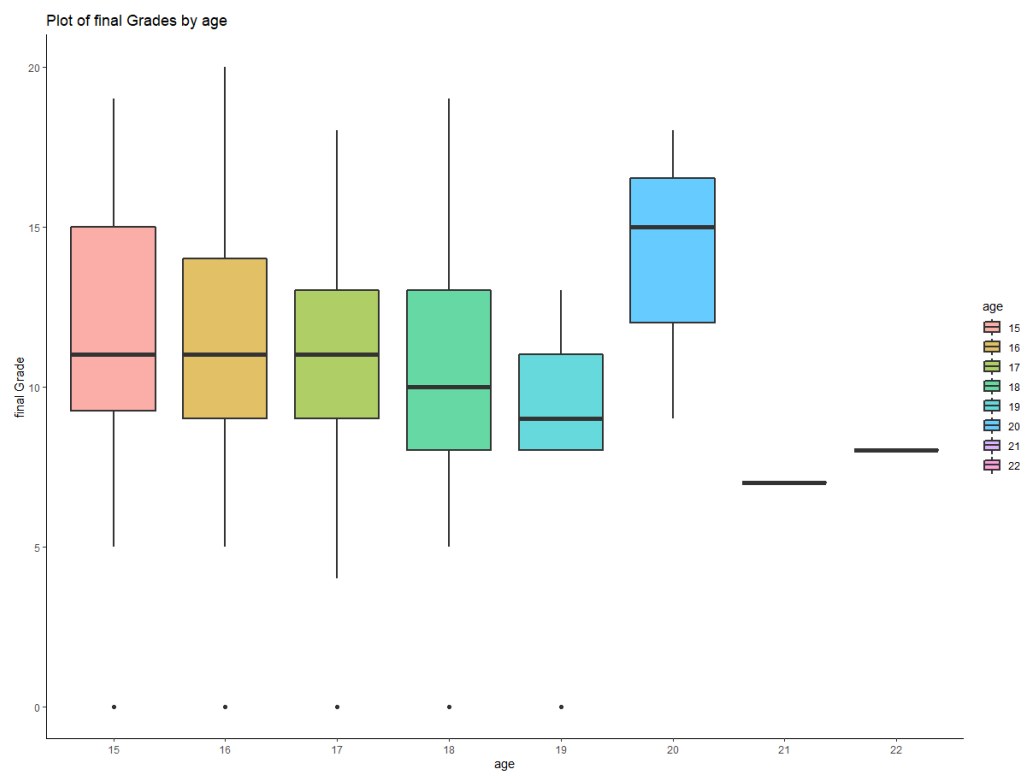
- Menampilkan beberapa grafik distribusi nilai akhir terhadap beberapa variabel

```
> corrplot(cor(studentDataSet[,unlist(lapply(studentDataSet, is.numeric))]))
> ggplot(studentDataSet, aes(x = sex, y = G3)) +
+   geom_boxplot(aes(fill = sex), alpha = .6, size = 1) +
+   ggtitle("Grade distribution by gender") +
+   theme_classic()
> |
```



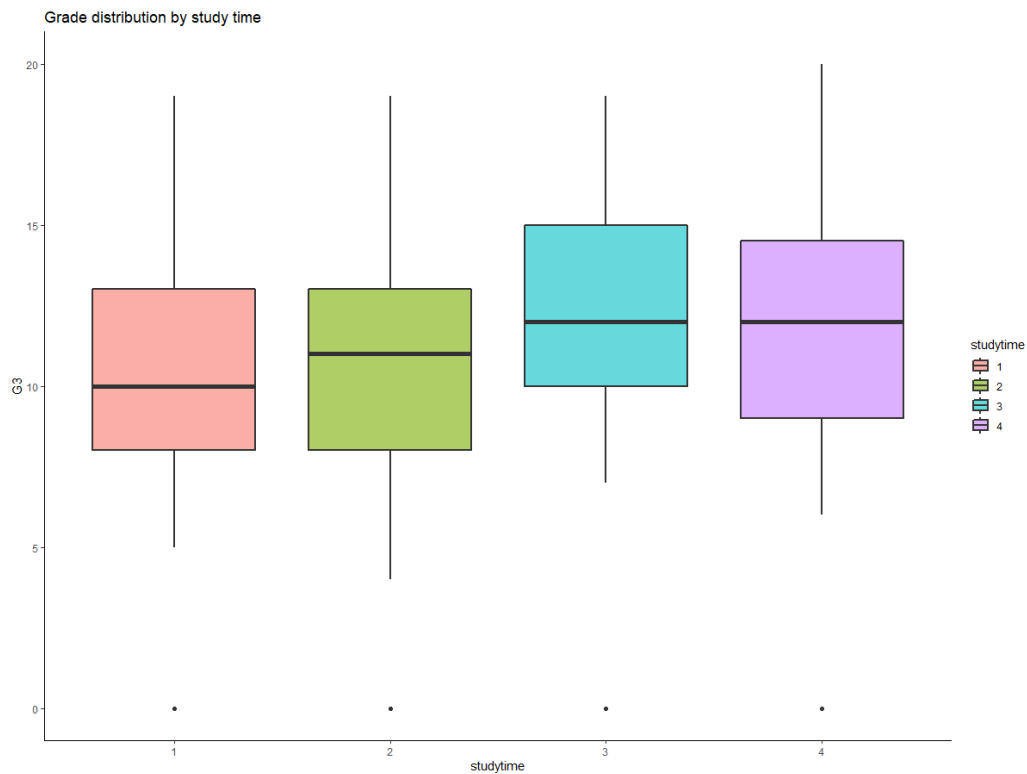
Berdasarkan gambar diatas, dapat disimpulkan bahwa nilai akhir anak laki-laki lebih unggul dibandingkan dengan nilai akhir anak perempuan

```
> studentDataSet$age = factor(studentDataSet$age)
> ggplot(studentDataSet, aes(x=age, y=G3, fill=age)) +
+   geom_boxplot(alpha = .6, size = 1) +
+   labs(title="Plot of final Grades by age", x="age", y = "final Grade") +
+   theme_classic()
> |
```



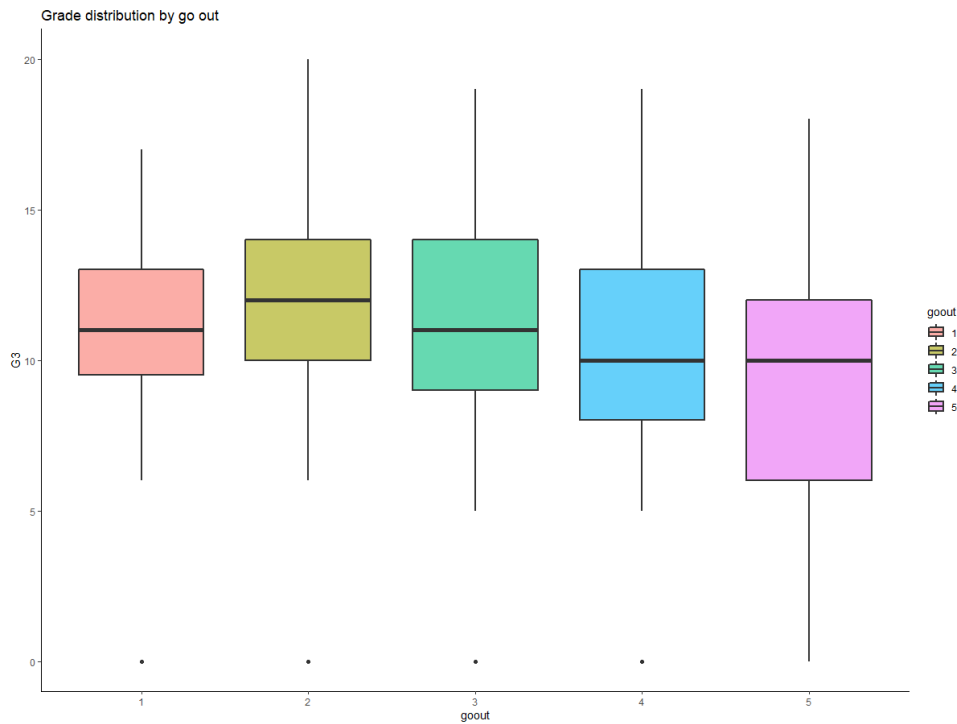
Berdasarkan gambar diatas, dapat disimpulkan bahwa anak berumur 20 mendapatkan hasil nilai akhir yang lebih tinggi dibandingkan dengan anak lainnya

```
> studentDataSet$studytime = factor(studentDataSet$studytime)
> ggplot(studentDataSet, aes(x = studytime, y = G3)) +
+   geom_boxplot(aes(fill = studytime),alpha = .6,size = 1) +
+   ggtitle("Grade distribution by study time") +
+   theme_classic()
> |
```



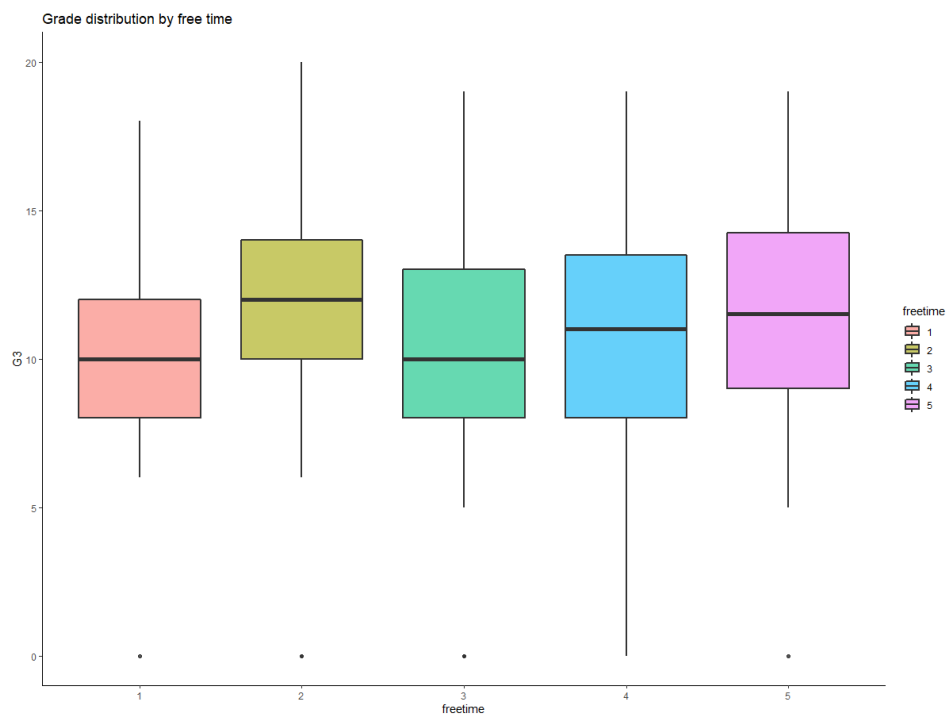
Berdasarkan gambar diatas, dapat disimpulkan bahwa anak dengan waktu belajar yang lebih banyak mendapatkan hasil nilai akhir yang lebih tinggi dibandingkan dengan anak lainnya

```
> studentDataSet$goout = factor(studentDataSet$goout)
> ggplot(studentDataSet, aes(x = goout, y = G3)) +
+   geom_boxplot(aes(fill = goout),alpha = .6,size = 1) +
+   ggtitle("Grade distribution by go out") +
+   theme_classic()
> |
```



Berdasarkan gambar diatas, dapat disimpulkan bahwa anak yang lebih jarang keluar untuk bermain dengan teman cenderung mendapatkan hasil nilai akhir yang lebih tinggi dibandingkan dengan anak lainnya

```
> studentDataSet$freetime = factor(studentDataSet$freetime)
> ggplot(studentDataSet, aes(x = freetime, y = G3)) +
+   geom_boxplot(aes(fill = freetime),alpha = .6,size = 1) +
+   ggtitle("Grade distribution by free time") +
+   theme_classic()
> |
```



Berdasarkan gambar diatas, dapat disimpulkan bahwa anak yang memiliki lebih banyak waktu luang cenderung mendapatkan hasil nilai akhir yang lebih tinggi dibandingkan dengan anak lainnya

- **Membuat model prediktif multiple regression**

Pertama-tama, kita import library yang akan digunakan untuk melakukan multiple regression

```
> library(caret)
> studentDataSet <- read.csv("student-mat.csv")
```

Selanjutnya, kita pisahkan data kita menjadi 2 yaitu, data train dan data test

```
> set.seed(1234)
> partition = createDataPartition(studentDataSet[, 'G3'], times=1, p=0.70, list=FALSE)
> training = studentDataSet[partition,]
> dim(training)
[1] 279 33
> test = studentDataSet[-partition,]
> dim(test)
[1] 116 33
> |
```

Berdasarkan hasil tersebut dapat dilihat bahwa data train memiliki 279 data, sedangkan data test memiliki 116 data.

Berikutnya, kita buat model regresinya

```
> lin_mod=lm(G3~G1+G2+age+Fedu+Medu+Fjob+Mjob+famsize+sex+Pstatus+absences+famsize+Dalc+famrel+traveltime, data=training)
> lin_mod
```

```
Call:
lm(formula = G3 ~ G1 + G2 + age + Fedu + Medu + Fjob + Mjob +
    famsize + sex + Pstatus + absences + famsize + Dalc + famrel +
    traveltime, data = training)
```

```
Coefficients:
(Intercept)          G1          G2          age          Fedu          Medu  Fjobhealth  Fjobother  Fjobservices  Fjobteacher
-0.04068      0.09746      1.00560     -0.13352     -0.09355      0.08506      0.15647     -0.28994     -0.95763     -0.15632
Mjobhealth  Mjobother  Mjobservices  Mjobteacher  famsizeLE3      sexM      PstatusT  absences      Dalc      famrel
-0.02662      0.41330      0.29147      0.43699     -0.15267      0.17302     -0.37891      0.03369     -0.01059      0.21494
traveltime
0.20147
```

Lalu kita juga dapat melihat ringkasan dari model regresi yang telah dibuat

```
> summary(lin_mod)

Call:
lm(formula = G3 ~ G1 + G2 + age + Fedu + Medu + Fjob + Mjob +
    famsize + sex + Pstatus + absences + famsize + Dalc + famrel +
    traveltime, data = training)

Residuals:
    Min       1Q   Median       3Q      Max
-8.8660 -0.4597  0.3594  0.9003  3.8140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.04068    1.91399   -0.021  0.9831
G1             0.09746    0.06926    1.407  0.1606
G2             1.00560    0.06122   16.425 <2e-16 ***
age           -0.13352    0.09546   -1.399  0.1631
Fedu          -0.09355    0.14214   -0.658  0.5110
Medu           0.08506    0.16505    0.515  0.6067
Fjobhealth     0.15647    0.71142    0.220  0.8261
Fjobother    -0.28994    0.52202   -0.555  0.5791
Fjobservices -0.95763    0.54366   -1.761  0.0793 .
Fjobteacher  -0.15632    0.66923   -0.234  0.8155
Mjobhealth   -0.02662    0.57834   -0.046  0.9633
Mjobother     0.41330    0.38109    1.085  0.2791
Mjobservices  0.29147    0.41149    0.708  0.4794
Mjobteacher   0.43699    0.53731    0.813  0.4168
famsizeLE3   -0.15267    0.25993   -0.587  0.5575
sexM           0.17302    0.23826    0.726  0.4684
PstatusT     -0.37891    0.41265   -0.918  0.3594
absences      0.03369    0.01652    2.040  0.0424 *
Dalc          -0.01059    0.13382   -0.079  0.9370
famrel        0.21494    0.13125    1.638  0.1027
traveltime    0.20147    0.16400    1.228  0.2204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.852 on 258 degrees of freedom
Multiple R-squared:  0.8357,    Adjusted R-squared:  0.8229
F-statistic: 65.6 on 20 and 258 DF,  p-value: < 2.2e-16
```

Selanjutnya kita dapat membuat prediksi terhadap G3 berdasarkan multiple regression model yang telah kita buat dengan menggunakan fungsi **predict()**

```
> pred_test <- predict(lin_mod, test)
> pred_test
      1      3      4      9     10     13     14     23     24     37     41
5.34076123 7.23040045 13.18977530 18.75715500 15.46888490 13.14051066 10.52221455 15.42843347 13.36694944 16.02126277 10.01535536
      42      44      47      48      56      57      58      65      67      69      76
12.47219413 7.15747777 11.26473721 18.83586387 9.38803196 14.84892976 15.72892005 8.88893990 12.38915398 7.57319361 9.12692342
      80      81      82      86      89      91      94      102      111      115      118
3.57540518 11.04055412 10.29177750 8.05369881 10.38601995 6.31396454 9.67161208 17.42338442 20.55186551 8.30551523 13.85963429
      119      120      123      124      136      137      140      141      143      145      146
7.66994466 13.22995068 13.31626997 11.13466534 -0.04221345 -0.29289345 16.72527663 7.84434160 10.07968742 -0.56364930 9.91661655
      149      151      152      156      157      158      164      169      171      176      177
5.40468163 3.53510908 12.72488174 6.91727390 11.84778608 7.67112137 8.69383889 6.43993474 4.46218335 8.91756913 12.87285513
      183      187      191      192      193      196      198      203      206      213      220
16.34976099 10.89114629 11.19411033 6.43442429 6.77547831 13.78662460 9.00044527 8.53828591 8.23163640 13.00729076 9.41203693
      222      224      225      231      232      234      244      246      251      253      259
3.67546031 12.69345787 12.49475548 12.54796696 11.11039535 12.81115915 11.51666195 19.32770071 7.29066704 7.35541637 14.62528585
      261      267      277      279      281      283      285      290      292      295      296
18.53885758 8.88598396 10.88353226 7.32090695 9.03468378 11.46329128 8.84220334 13.92641410 13.98233698 13.44070906 11.75849663
      302      303      305      307      310      312      316      321      322      323      326
10.92902367 12.36817160 14.57907243 18.55461055 9.47490922 12.17412245 11.78309939 13.46306179 8.99424499 9.53449936 11.52482966
      329      333      334      337      338      339      344      353      354      373      375
8.28230575 -1.56682299 7.00578625 13.98253904 7.26899505 14.47740950 7.43005209 5.13885801 7.72295513 11.14059197 18.53323470
      386      388      390      391      392      393
8.42946185 3.91025376 3.51187247 8.15798436 15.18241834 7.42914832
```

Data diatas merupakan hasil prediksi terhadap nilai G3

Discussion and Analysis

Kita lakukan analisa terhadap multiple regression model yang telah dibuat

```
> lin_mod=lm(G3~G1+G2+age+Fedu+Medu+Fjob+Mjob+famsize+sex+Pstatus+absences+famsize+Dalc+famrel+traveltime, data=training)
> lin_mod

Call:
lm(formula = G3 ~ G1 + G2 + age + Fedu + Medu + Fjob + Mjob +
    famsize + sex + Pstatus + absences + famsize + Dalc + famrel +
    traveltime, data = training)

Coefficients:
(Intercept)      G1      G2      age      Fedu      Medu  Fjobhealth  Fjobother  Fjobservices  Fjobteacher
-0.04068      0.09746    1.00560 -0.13352 -0.09355    0.08506    0.15647   -0.28994   -0.95763   -0.15632
Mjobhealth      Mjobother  Mjobservices  Mjobteacher  famsizeLE3      sexM      PstatusT      absences      Dalc      famrel
-0.02662      0.41330    0.29147    0.43699   -0.15267    0.17302   -0.37891    0.03369   -0.01059    0.21494
traveltime
0.20147

> summary(lin_mod)

Call:
lm(formula = G3 ~ G1 + G2 + age + Fedu + Medu + Fjob + Mjob +
    famsize + sex + Pstatus + absences + famsize + Dalc + famrel +
    traveltime, data = training)

Residuals:
    Min       1Q   Median       3Q      Max
-8.8660 -0.4597  0.3594  0.9003  3.8140

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04068     1.91399  -0.021  0.9831
G1           0.09746     0.06926   1.407  0.1606
G2           1.00560     0.06122  16.425 <2e-16 ***
age          -0.13352     0.09546  -1.399  0.1631
Fedu         -0.09355     0.14214  -0.658  0.5110
Medu         0.08506     0.16505   0.515  0.6067
Fjobhealth   0.15647     0.71142   0.220  0.8261
Fjobother    -0.28994     0.52202  -0.555  0.5791
Fjobservices -0.95763     0.54366  -1.761  0.0793 .
Fjobteacher  -0.15632     0.66923  -0.234  0.8155
Mjobhealth   -0.02662     0.57834  -0.046  0.9633
Mjobother     0.41330     0.38109   1.085  0.2791
Mjobservices  0.29147     0.41149   0.708  0.4794
Mjobteacher   0.43699     0.53731   0.813  0.4168
famsizeLE3   -0.15267     0.25993  -0.587  0.5575
sexM         0.17302     0.23826   0.726  0.4684
PstatusT     -0.37891     0.41265  -0.918  0.3594
absences      0.03369     0.01652   2.040  0.0424 *
Dalc         -0.01059     0.13382  -0.079  0.9370
famrel        0.21494     0.13125   1.638  0.1027
traveltime    0.20147     0.16400   1.228  0.2204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.852 on 258 degrees of freedom
Multiple R-squared:  0.8357,    Adjusted R-squared:  0.8229
F-statistic: 65.6 on 20 and 258 DF,  p-value: < 2.2e-16
```

Berdasarkan ringkasan tersebut, dapat dilihat bahwa multiple regression model tersebut memiliki standard error yang cukup kecil yang artinya multiple regression model ini memiliki tingkat error yang cukup kecil. Hasil R-squared yang ditunjukkan juga cukup besar yaitu 83,5% yang artinya multiple regression model ini telah mewakili keragaman dataset tersebut. F-statistik yang dihasilkan juga terbilang cukup besar yang artinya multiple regression model ini sangat baik dalam hal memprediksi.

Selanjutnya kita dapat membuat prediksi terhadap G3 berdasarkan multiple regression model yang telah kita buat

```
> pred_test <- predict(lin_mod, test)
> pred_test
      1      3      4      9     10     13     14     23     24     37     41
5.34076123 7.23040045 13.18977530 18.75715500 15.46888490 13.14051066 10.52221455 15.42843347 13.36694944 16.02126277 10.01535536
      42      44      47      48      56      57      58      65      67      69      76
12.47219413 7.15747777 11.26473721 18.83586387 9.38803196 14.84892976 15.72892005 8.88893990 12.38915398 7.57319361 9.12692342
      80      81      82      86      89      91      94      102      111      115      118
3.57540518 11.04055412 10.29177750 8.05369881 10.38601995 6.31396454 9.67161208 17.42338442 20.55186551 8.30551523 13.85963429
      119      120      123      124      136      137      140      141      143      145      146
7.66994466 13.22995068 13.31626997 11.13466534 -0.04221345 -0.29289345 16.72527663 7.84434160 10.07968742 -0.56364930 9.91661655
      149      151      152      156      157      158      164      169      171      176      177
5.40468163 3.53510908 12.72488174 6.91727390 11.84778608 7.67112137 8.69383889 6.43993474 4.46218335 8.91756913 12.87285513
      183      187      191      192      193      196      198      203      206      213      220
16.34976099 10.89114629 11.19411033 6.43442429 6.77547831 13.78662460 9.00044527 8.53828591 8.23163640 13.00729076 9.41203693
      222      224      225      231      232      234      244      246      251      253      259
3.67546031 12.69345787 12.49475548 12.54796696 11.11039535 12.81115915 11.51666195 19.32770071 7.29066704 7.35541637 14.62528585
      261      267      277      279      281      283      285      290      292      295      296
18.53857558 8.88598396 10.88353226 7.32090695 9.03468378 11.46329128 8.84220334 13.92641410 13.98233698 13.44070906 11.75849663
      302      303      305      307      310      312      316      321      322      323      326
10.92902367 12.36817160 14.57907243 18.55461055 9.47490922 12.17412245 11.78309939 13.46306179 8.99424499 9.53449936 11.52482966
      329      333      334      337      338      339      344      353      354      355      375
8.28230575 -1.56682299 7.00578625 13.98253904 7.26899505 14.47740950 7.43005209 5.13885801 7.72295513 11.14059197 18.53323470
      386      390      391      392      393
8.42946185 3.91025376 3.51187247 8.15798436 15.18241834 7.42914832
> |
```

Selanjutnya kita dapat mengevaluasi model yang kita buat

Metrik yang akan kita gunakan dalam evaluasi ini adalah MAE. Performa model kita bagus jika nilai MAE rendah. Jadi kita ingin mendapatkan nilai MAE yang rendah.

Mean Absolute Error (MAE) mengukur besarnya rata-rata kesalahan dalam satu set prediksi, tanpa mempertimbangkan arahnya. Ini adalah rata-rata atas sampel uji dari perbedaan mutlak antara prediksi dan pengamatan aktual di mana semua perbedaan individu memiliki bobot yang sama.

```
> cat("Test MAE:", round(mean(abs(pred_test-test$G3)),6))
Test MAE: 1.395067
> |
```

Lalu, mari kita cek apakah model kita overfitted dengan menghitung MAE train dan membandingkannya dengan MAE tes

```
> pred_train <- predict(lin_mod, training)
> cat("Train MAE:", round(mean(abs(pred_train-training$G3)),6))
Train MAE: 1.112655
> |
```

Nilai MAE dari data train dan data test nya menunjukkan angka yang hampir sama, sehingga model regresi ini tidak overfit.

Namun nilai MAE menunjukkan angka yang kecil, sehingga model regresi yang dibuat berjalan dengan baik

Terakhir kita dapat membandingkan hasil prediksi nilai G3 dengan hasil G3 yang sebenarnya

```
> G3prediction = (data.frame((pred_test), (test$G3)))
> colnames(G3prediction) <- c("Predicted G3", "Real G3")
> head(G3prediction,10)
      Predicted G3 Real G3
1          5.340761      6
3          7.230400     10
4         13.189775     15
9         18.757155     19
10        15.468885     15
13        13.140511     14
14        10.522215     11
23        15.428433     16
24        13.366949     12
37        16.021263     18
> |
```

Dengan membandingkan hasil prediksi nilai G3 dengan hasil G3 sebenarnya, dapat dilihat bahwa prediksi yang ditunjukkan hampir mendekati angka sebenarnya dengan presentase error yang sangat kecil, sehingga multiple regression modelnya dapat dikatakan cukup baik.

Kesimpulan :

Berdasarkan hasil tes korelasi variabel numerik, dapat disimpulkan bahwa hanya beberapa variabel saja yang mempengaruhi hasil nilai akhir siswa seperti hasil nilai ujian pertama dan hasil nilai ujian kedua. Namun variabel numerik lainnya hampir tidak memiliki hubungan yang kuat dengan hasil nilai akhir siswa. Lalu, berdasarkan grafik visualisasi, dapat dilihat bahwa siswa yang memiliki jam belajar yang lebih banyak, memiliki waktu luang yang lebih banyak, serta lebih jarang keluar berpergian dengan teman cenderung memiliki nilai akhir yang lebih tinggi dibandingkan dengan siswa lainnya. Berdasarkan prediksi multiple regression model yang telah dibuat, model regresinya cukup mewakili keragaman dataset dan dapat berjalan dengan baik karena hasil prediksi G3 menunjukkan angka yang hampir mendekati angka G3 yang sebenarnya.