EDA_Portugues_Students

Gabe Vasquez

2025-05-11

Summary

There are 649 total students with 33 variables.
The response variable is GPA (continuous).
There are 16 numerical variables and 17 categorical variables.

Our variables of interest are:
-Numerical variables: G1, G2, and Failures
-Categorical variables: Absences, Medu, Fedu, Studytime

Others that showed potential from the their figures: School, Sex, Address, Mjob, Fjob, Guardian, Higher, Schoolsup, Internet.

From the scatterplots:

From the boxplots:

From the EDA, our desired model appears to be:

$$GPA = B_0 + B_1(G1) + B_2(G2) + B_3() + B_4() + B_5() + B_6() + B_6() + B_8()$$

```
# Load the data from CSV files
port_students <- read.csv("student-por.csv", sep=";", header=TRUE)
```

```
# Generate GPA. Round GPA to 2 decimal place.
port_students <- port_students %>%
  mutate(GPA = (G3 / 20) * 4)

port_students <- port_students %>%
  mutate(GPA = round(GPA, 2))

head(port_students)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 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
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3 GPA
## 1        4  0 11 11 2.2
## 2        2  9 11 11 2.2
## 3        6 12 13 12 2.4
## 4        0 14 14 14 2.8
## 5        0 11 13 13 2.6
## 6        6 12 12 13 2.6
```

Portuguese Language Students

General Data Structure and Summary

```
# General structure and summary
str(port_students)
```

```
## 'data.frame':    649 obs. of  34 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "no" "no" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : int  0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : int  11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
##  $ GPA       : num  2.2 2.2 2.4 2.8 2.6 2.6 2.6 2.6 3.4 2.6 ...
```

```
glimpse(port_students)
```

```
## Rows: 649
## Columns: 34
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP",…
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",…
## $ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,…
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",…
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE…
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T",…
## $ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,…
## $ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3,…
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "servic…
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other", …
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation", …
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mother",…
## $ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,…
## $ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,…
## $ failures   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,…
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", "n…
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes",…
## $ paid       <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no",…
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye…
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "yes…
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye…
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",…
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "no"…
## $ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,…
## $ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,…
## $ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,…
## $ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,…
## $ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,…
## $ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,…
## $ absences   <int> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2, 2, 6…
## $ G1         <int> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12, 14, 1…
## $ G2         <int> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, 12, 14,…
## $ G3         <int> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13, 15,…
## $ GPA        <dbl> 2.2, 2.2, 2.4, 2.8, 2.6, 2.6, 2.6, 2.6, 3.4, 2.6, 2.8, 2.6,…
```

```
skimr::skim(port_students)
```

Data summary

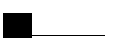
Name	port_students
Number of rows	649
Number of columns	34
Column type frequency:	
character	17

numeric	17
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
school	0	1	2	2	0	2	0
sex	0	1	1	1	0	2	0
address	0	1	1	1	0	2	0
famsize	0	1	3	3	0	2	0
Pstatus	0	1	1	1	0	2	0
Mjob	0	1	5	8	0	5	0
Fjob	0	1	5	8	0	5	0
reason	0	1	4	10	0	4	0
guardian	0	1	5	6	0	3	0
schoolsup	0	1	2	3	0	2	0
famsup	0	1	2	3	0	2	0
paid	0	1	2	3	0	2	0
activities	0	1	2	3	0	2	0
nursery	0	1	2	3	0	2	0
higher	0	1	2	3	0	2	0
internet	0	1	2	3	0	2	0
romantic	0	1	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	16.74	1.22	15	16	17.0	18.0	22.0	
Medu	0	1	2.51	1.13	0	2	2.0	4.0	4.0	
Fedu	0	1	2.31	1.10	0	1	2.0	3.0	4.0	
traveltime	0	1	1.57	0.75	1	1	1.0	2.0	4.0	
studytime	0	1	1.93	0.83	1	1	2.0	2.0	4.0	
failures	0	1	0.22	0.59	0	0	0.0	0.0	3.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
famrel	0	1	3.93	0.96	1	4	4.0	5.0	5.0	
freetime	0	1	3.18	1.05	1	3	3.0	4.0	5.0	
goout	0	1	3.18	1.18	1	2	3.0	4.0	5.0	
Dalc	0	1	1.50	0.92	1	1	1.0	2.0	5.0	
Walc	0	1	2.28	1.28	1	1	2.0	3.0	5.0	
health	0	1	3.54	1.45	1	2	4.0	5.0	5.0	
absences	0	1	3.66	4.64	0	0	2.0	6.0	32.0	
G1	0	1	11.40	2.75	0	10	11.0	13.0	19.0	
G2	0	1	11.57	2.91	0	10	11.0	13.0	19.0	
G3	0	1	11.91	3.23	0	10	12.0	14.0	19.0	
GPA	0	1	2.38	0.65	0	2	2.4	2.8	3.8	

```
# Check for missing values
missing_counts <- colSums(is.na(port_students))
print(missing_counts[missing_counts > 0])
```

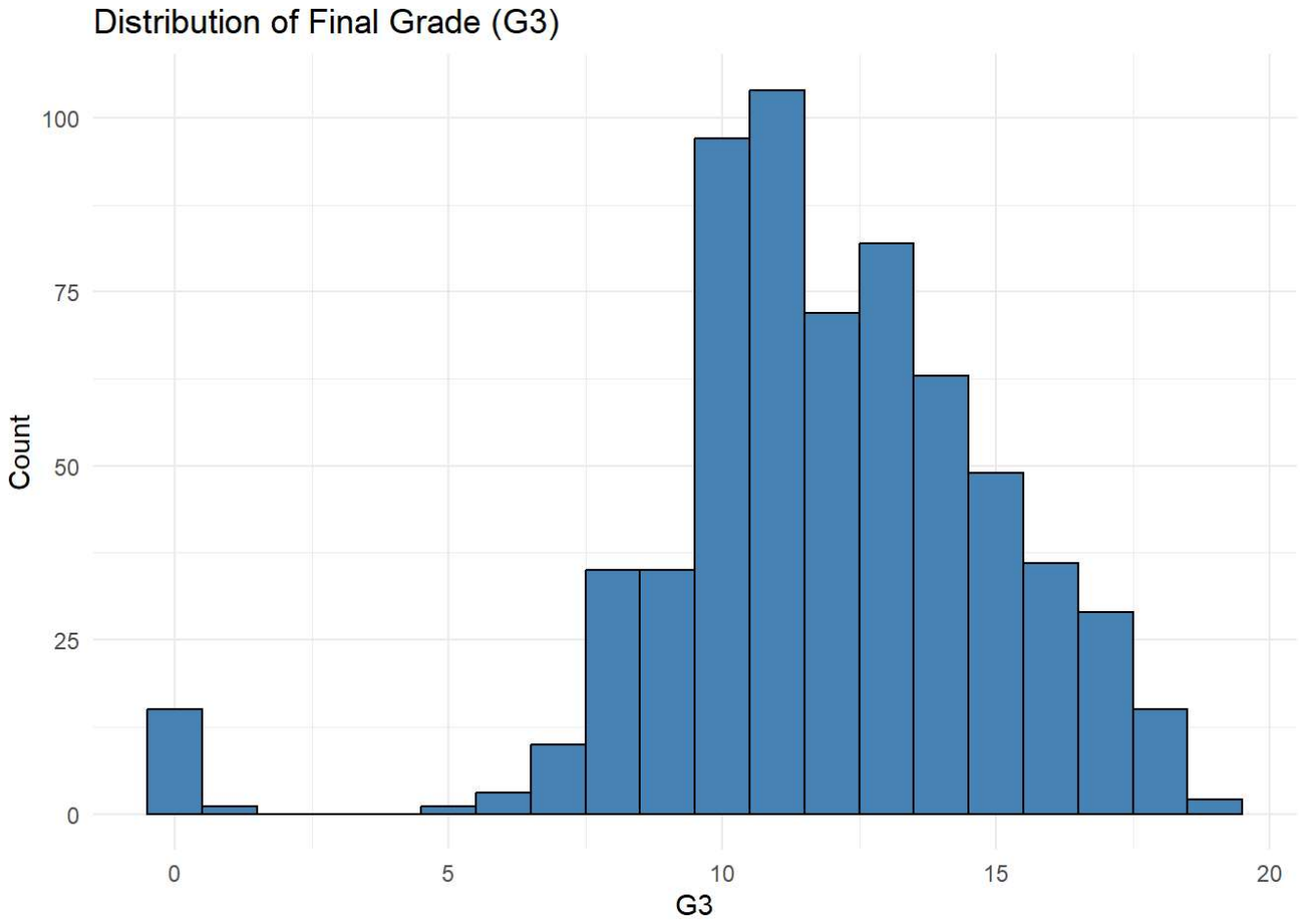
```
## named numeric(0)
```

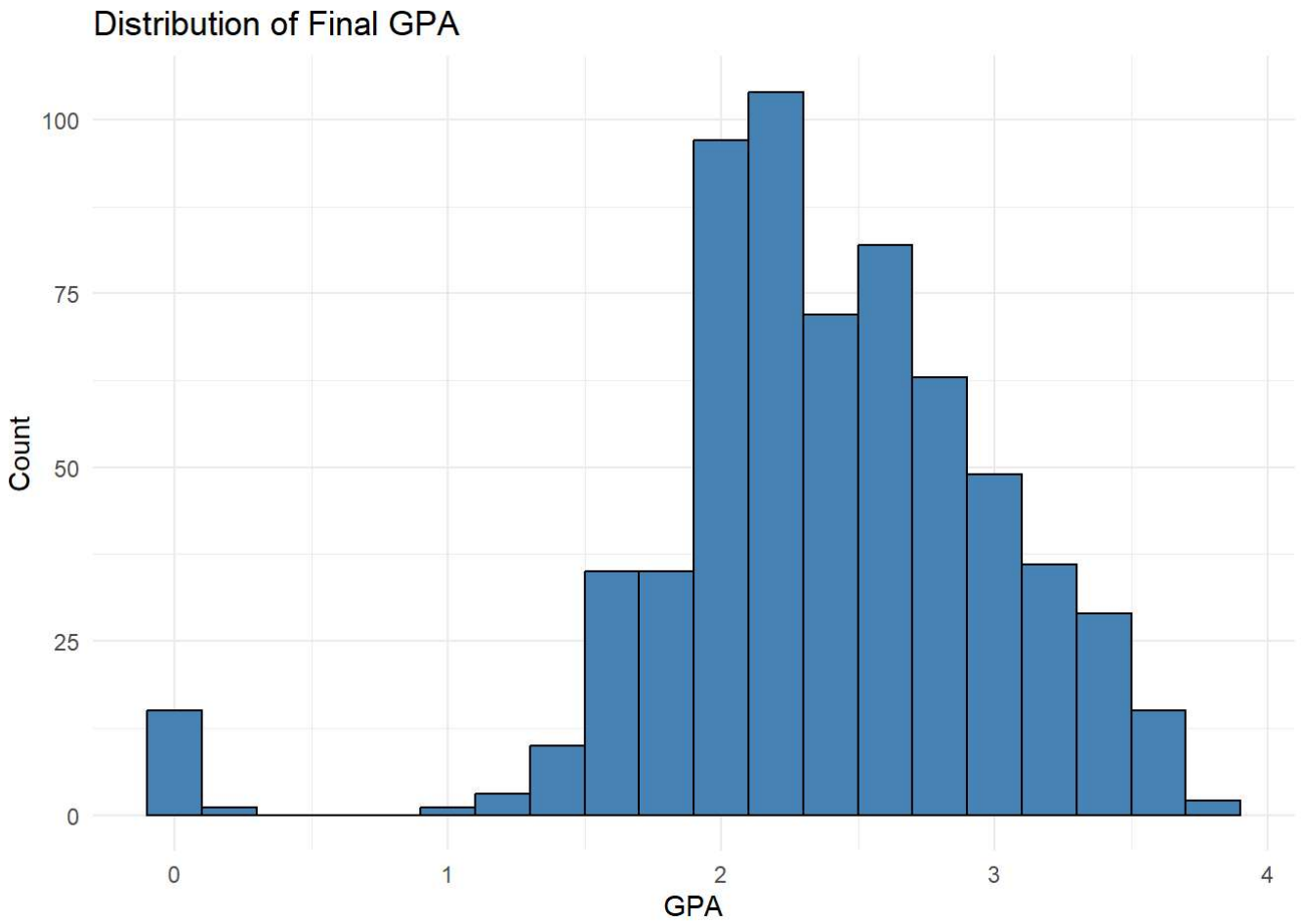
There are no missing values in this dataset.

```
# Summary statistics for numeric variables
summary(select(port_students, where(is.numeric)))
```

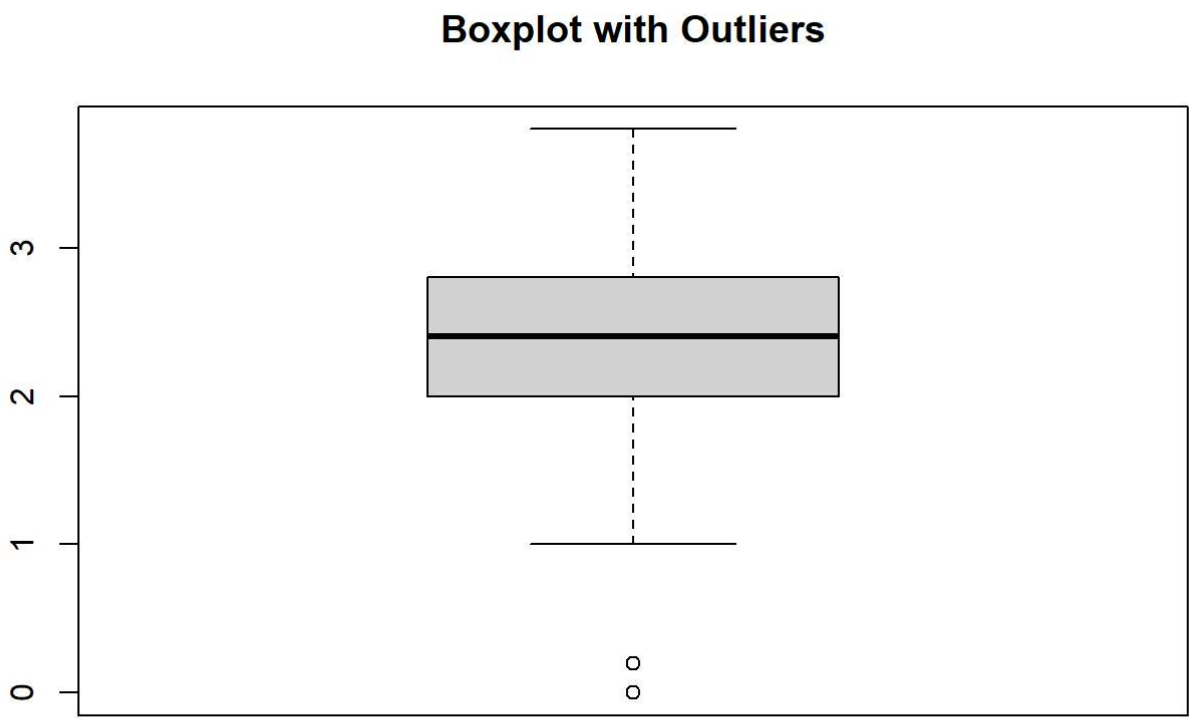
```
##       age             Medu            Fedu         traveltime
##  Min.   :15.00   Min.   :0.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:16.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :17.00   Median :2.000   Median :2.000   Median :1.000
##  Mean   :16.74   Mean   :2.515   Mean   :2.307   Mean   :1.569
##  3rd Qu.:18.00   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :22.00   Max.   :4.000   Max.   :4.000   Max.   :4.000
##    studytime        failures          famrel         freetime
##  Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:4.000   1st Qu.:3.00
##  Median :2.000   Median :0.0000   Median :4.000   Median :3.00
##  Mean   :1.931   Mean   :0.2219   Mean   :3.931   Mean   :3.18
##  3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.:4.00
##  Max.   :4.000   Max.   :3.0000   Max.   :5.000   Max.   :5.00
##     goout            Dalc            Walc          health
##  Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000
##  Median :3.000   Median :1.000   Median :2.00   Median :4.000
##  Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000
##  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
##       GPA
##  Min.   :0.000
##  1st Qu.:2.000
##  Median :2.400
##  Mean   :2.381
##  3rd Qu.:2.800
##  Max.   :3.800
```

Distribution of Response Variable






```
# Check for outliers
boxplot(port_students$GPA, main="Boxplot with Outliers")
```




```
iqr <- IQR(port_students$GPA)
lower <- quantile(port_students$GPA, 0.25) - 1.5 * iqr
upper <- quantile(port_students$GPA, 0.75) + 1.5 * iqr
outliers <- port_students$GPA[port_students$GPA < lower | port_students$GPA > upper]

print(outliers)
```

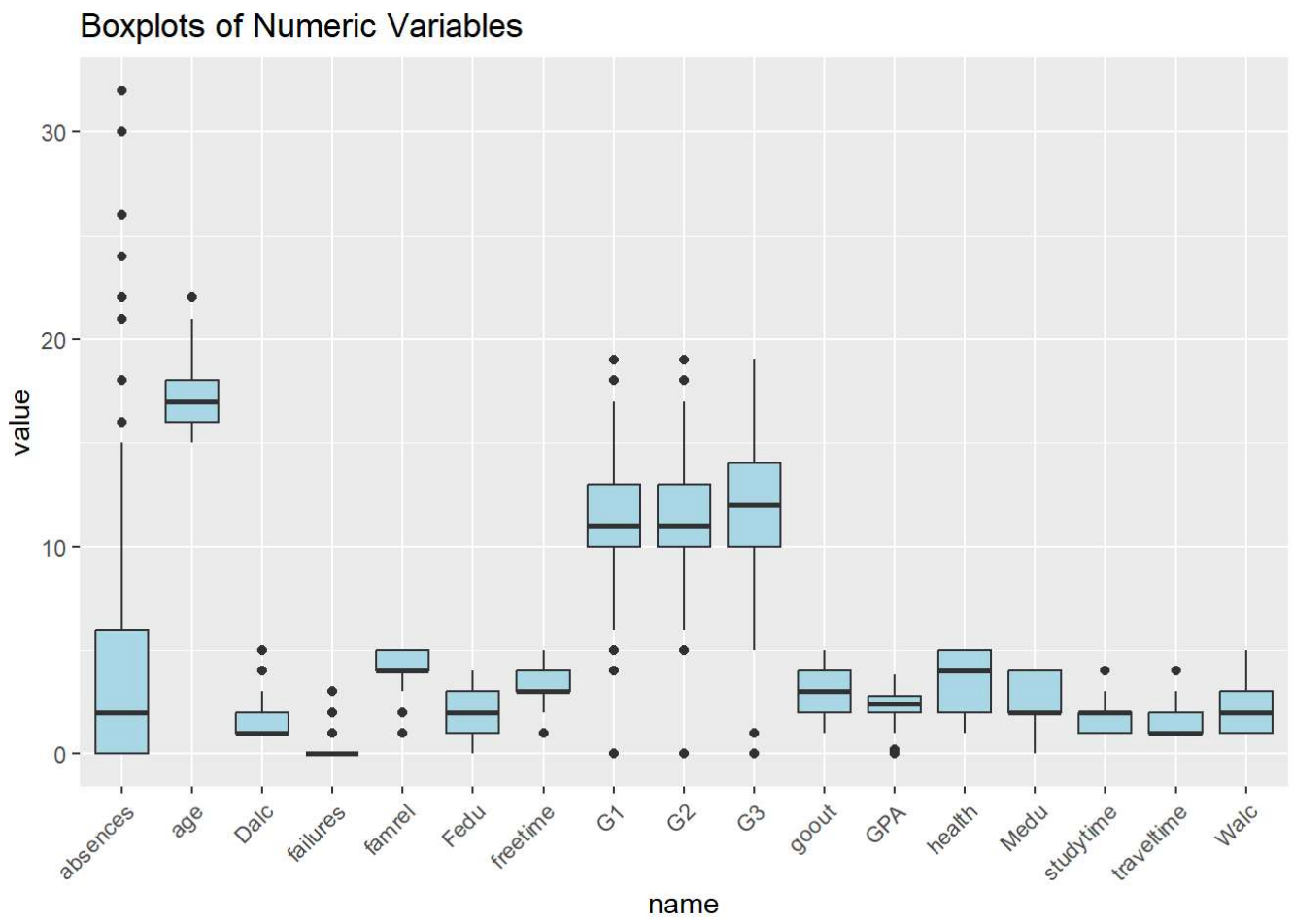
```
##  [1] 0.0 0.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

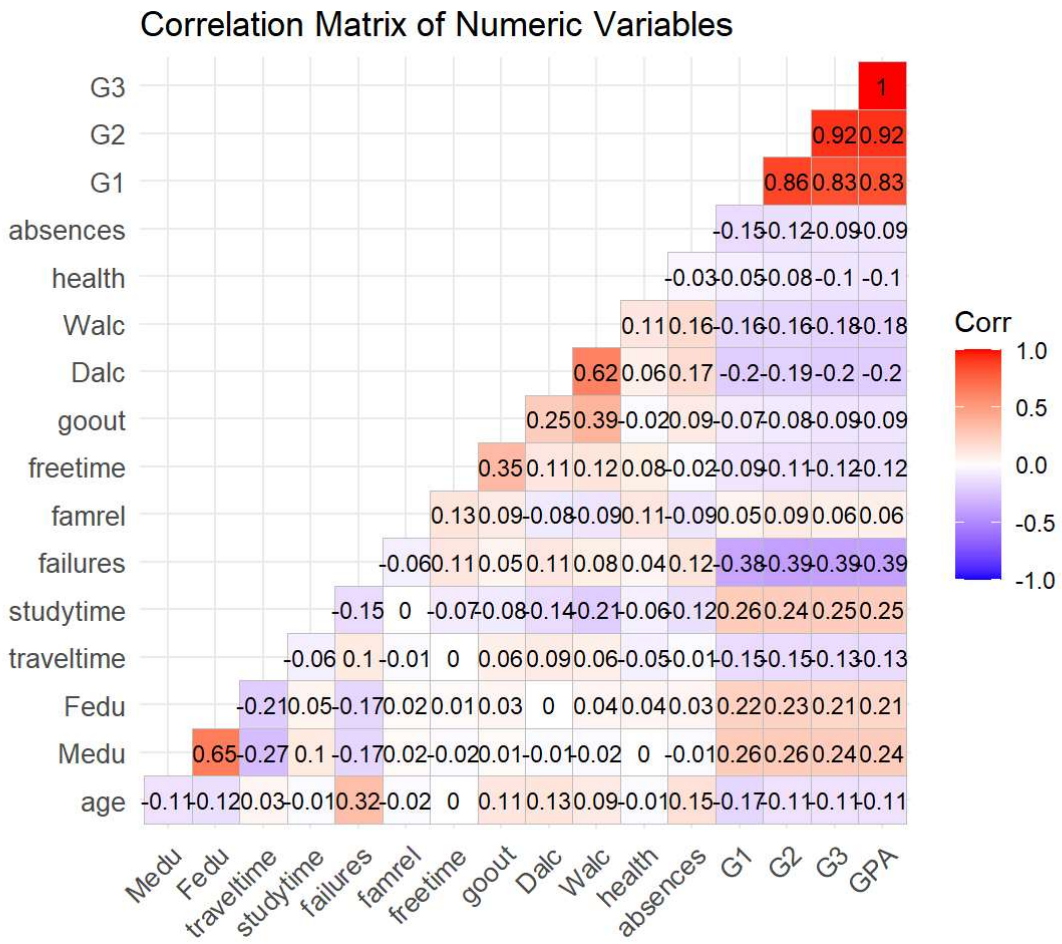
```
length(outliers)
```

```
## [1] 16
```

There are 16 extreme outliers in GPA. GPA also does not follow the traditional normal distribution. In linear regression approach the idea would be to expect to remove these 16 outliers from the dataset. However, in our quantile regression approach, let's see if we gain insights into these extreme students with very low GPA scores.

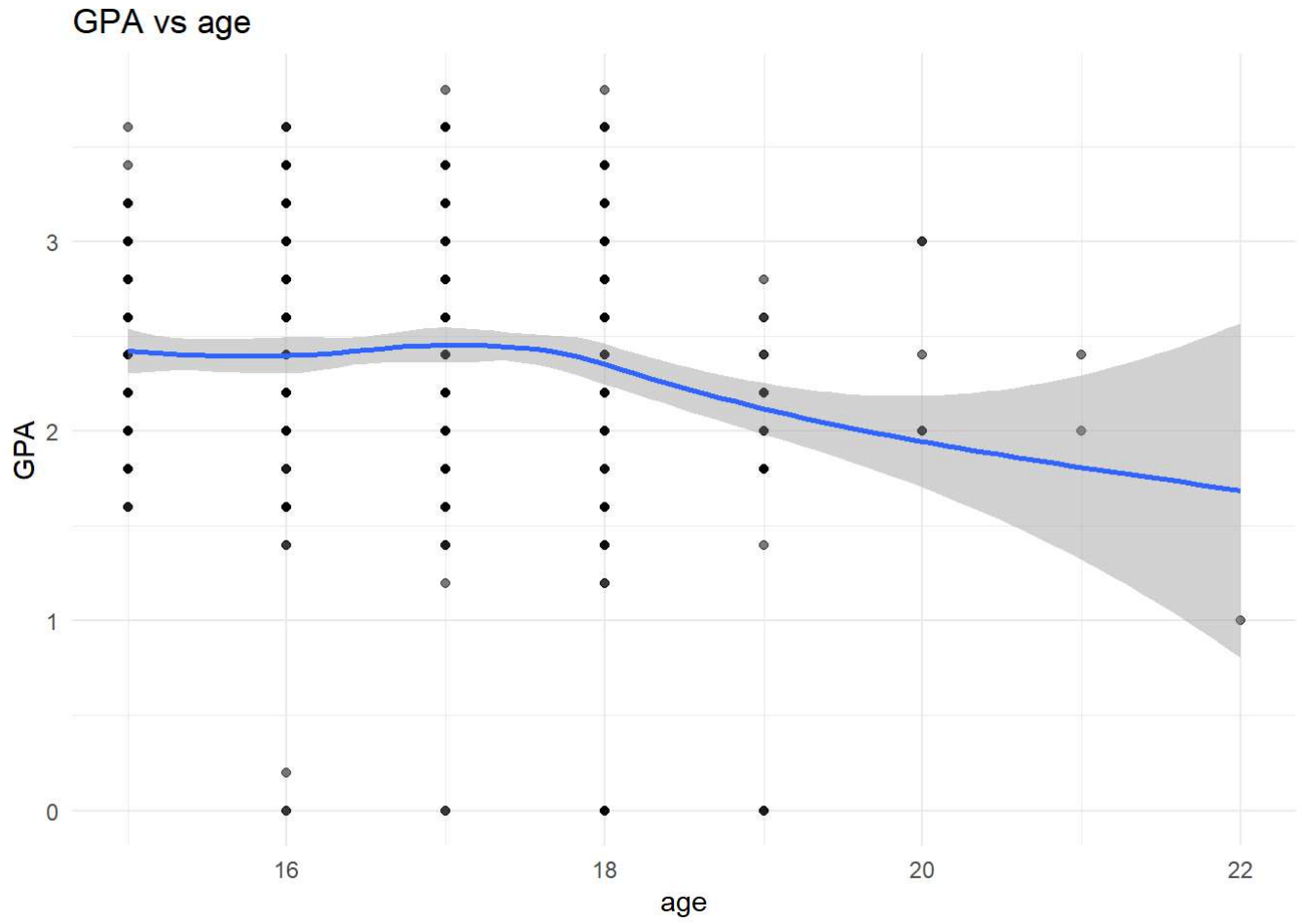
Correlations



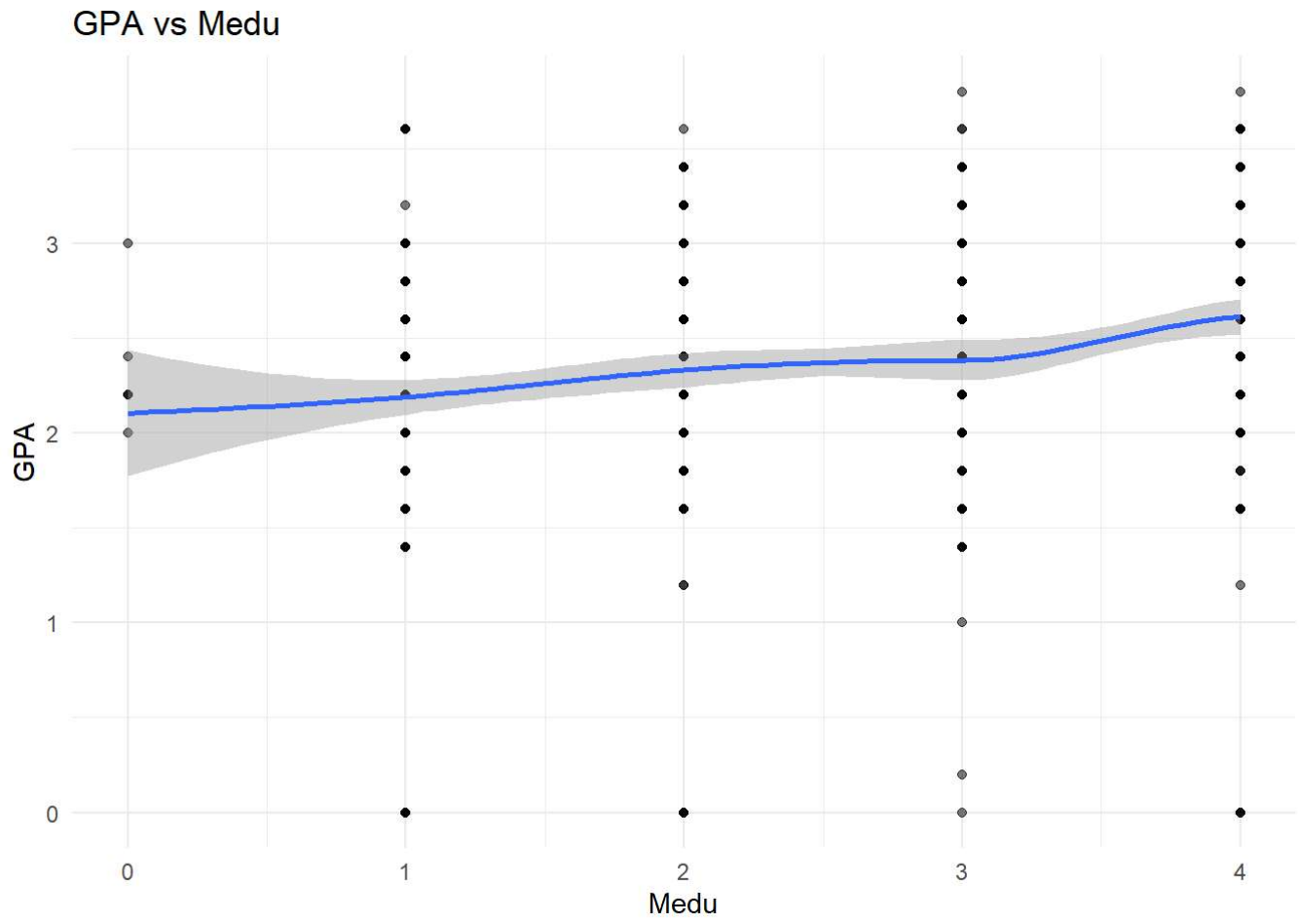


Relationships with GPA

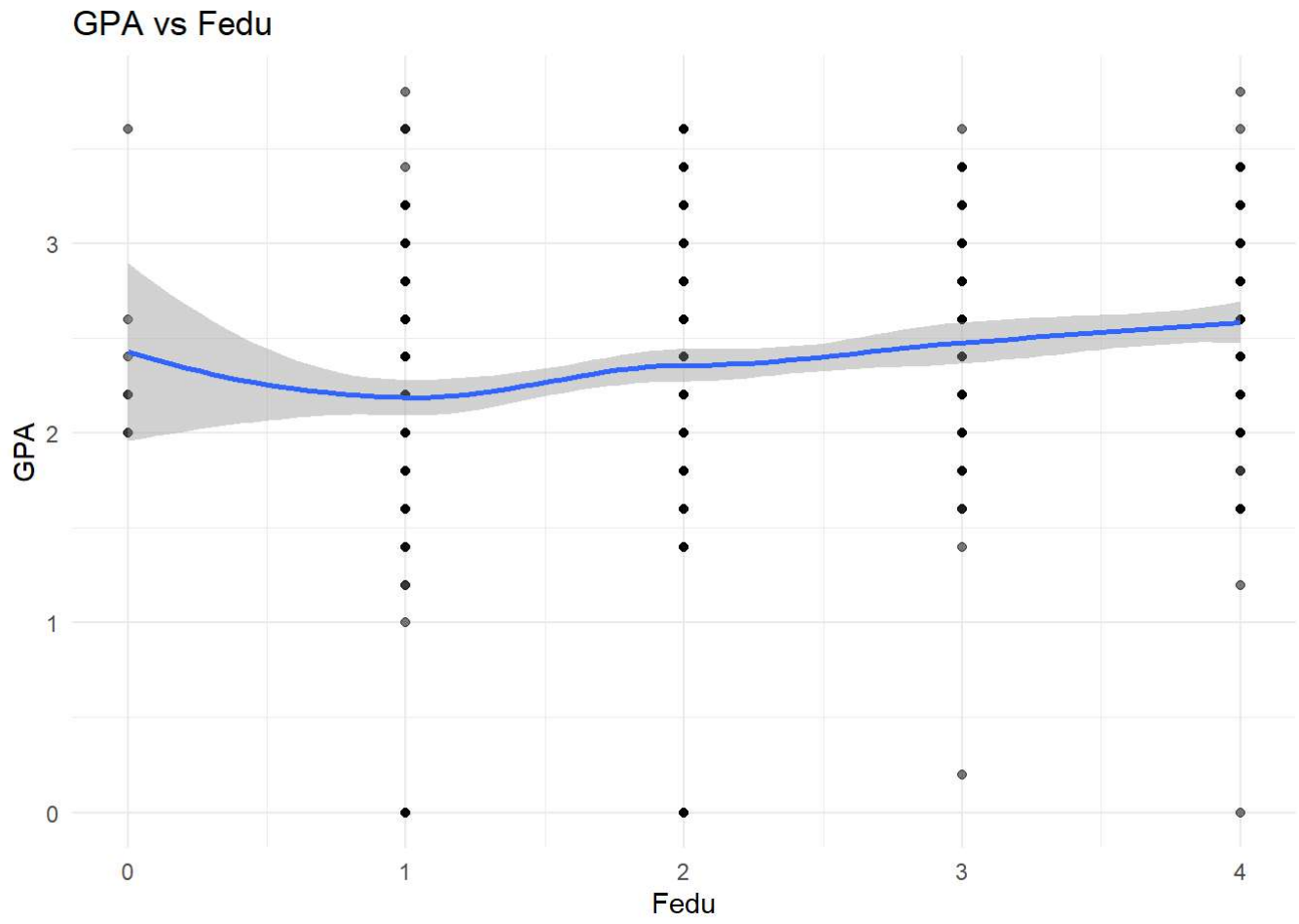
```
## `geom_smooth()` using formula = 'y ~ x'
```



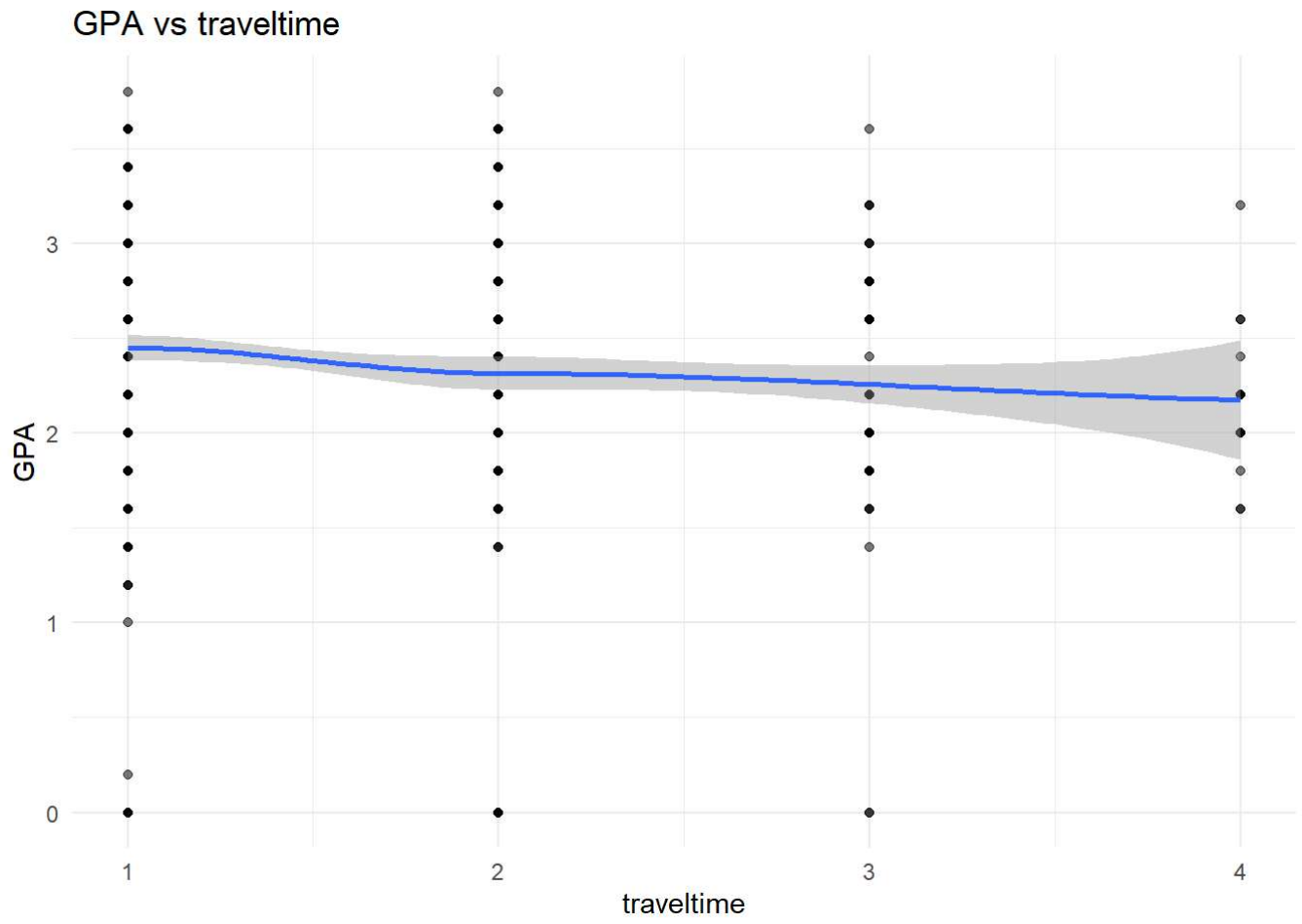

```
## `geom_smooth()` using formula = 'y ~ x'
```



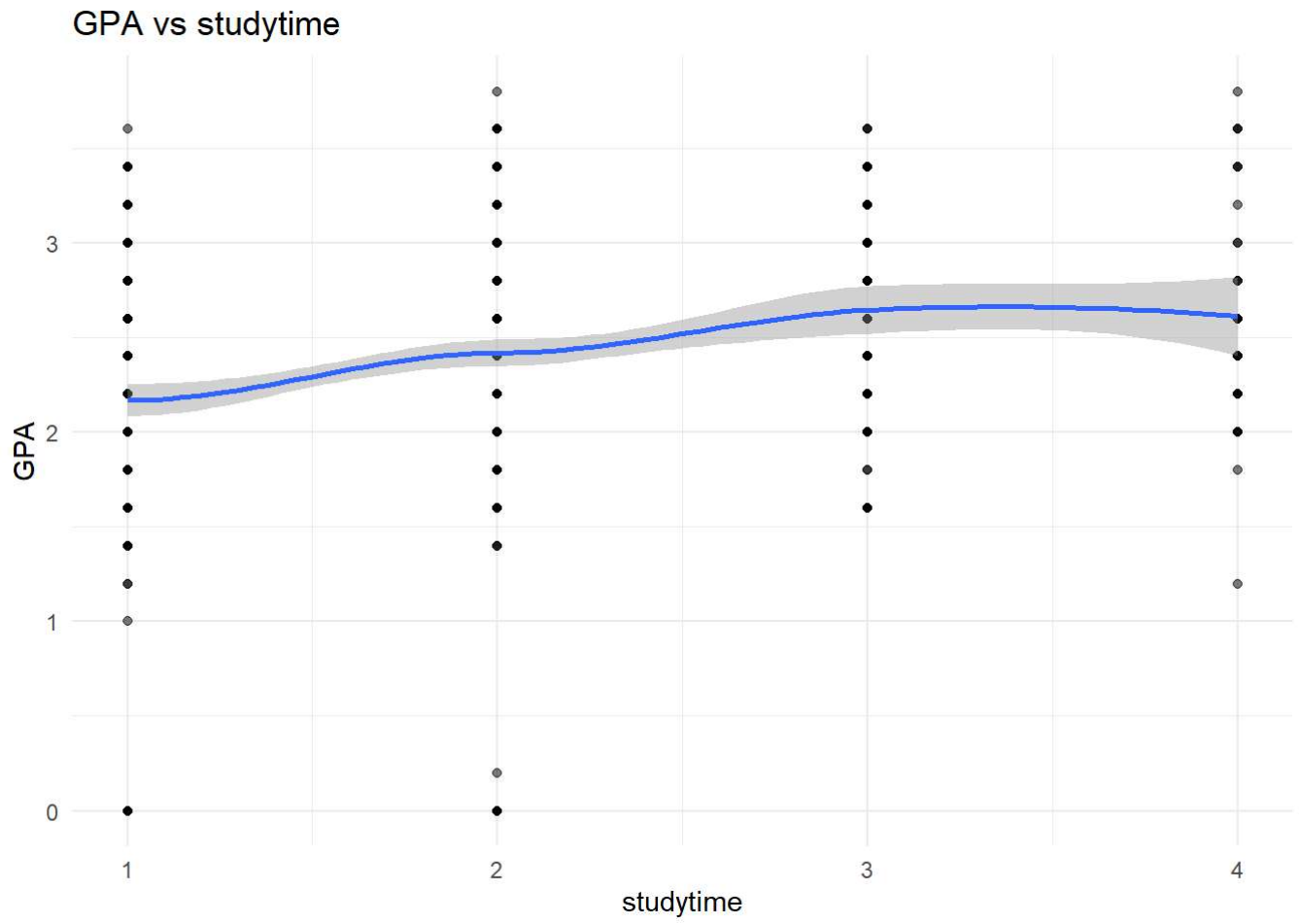

```
## `geom_smooth()` using formula = 'y ~ x'
```



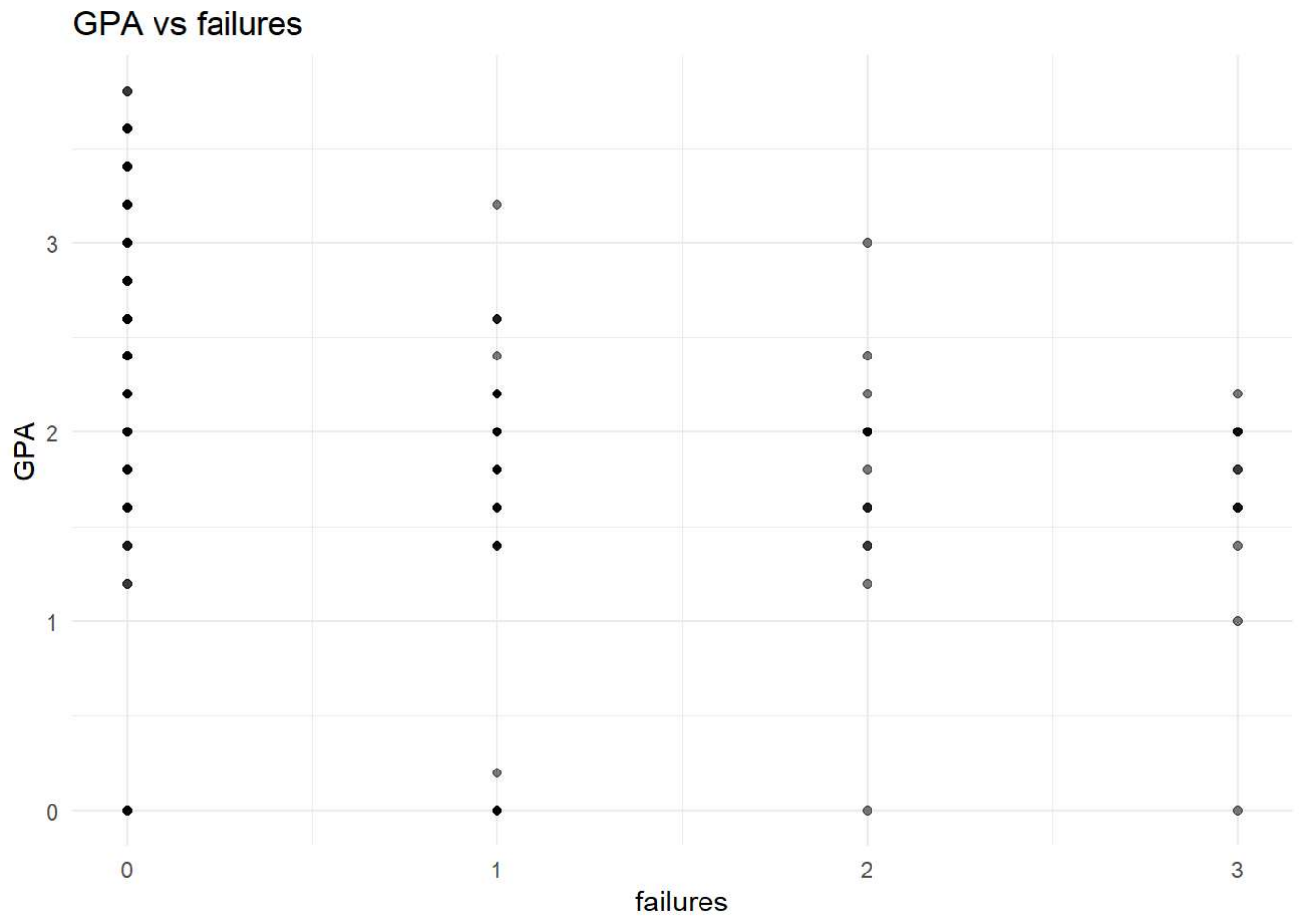

```
## `geom_smooth()` using formula = 'y ~ x'
```



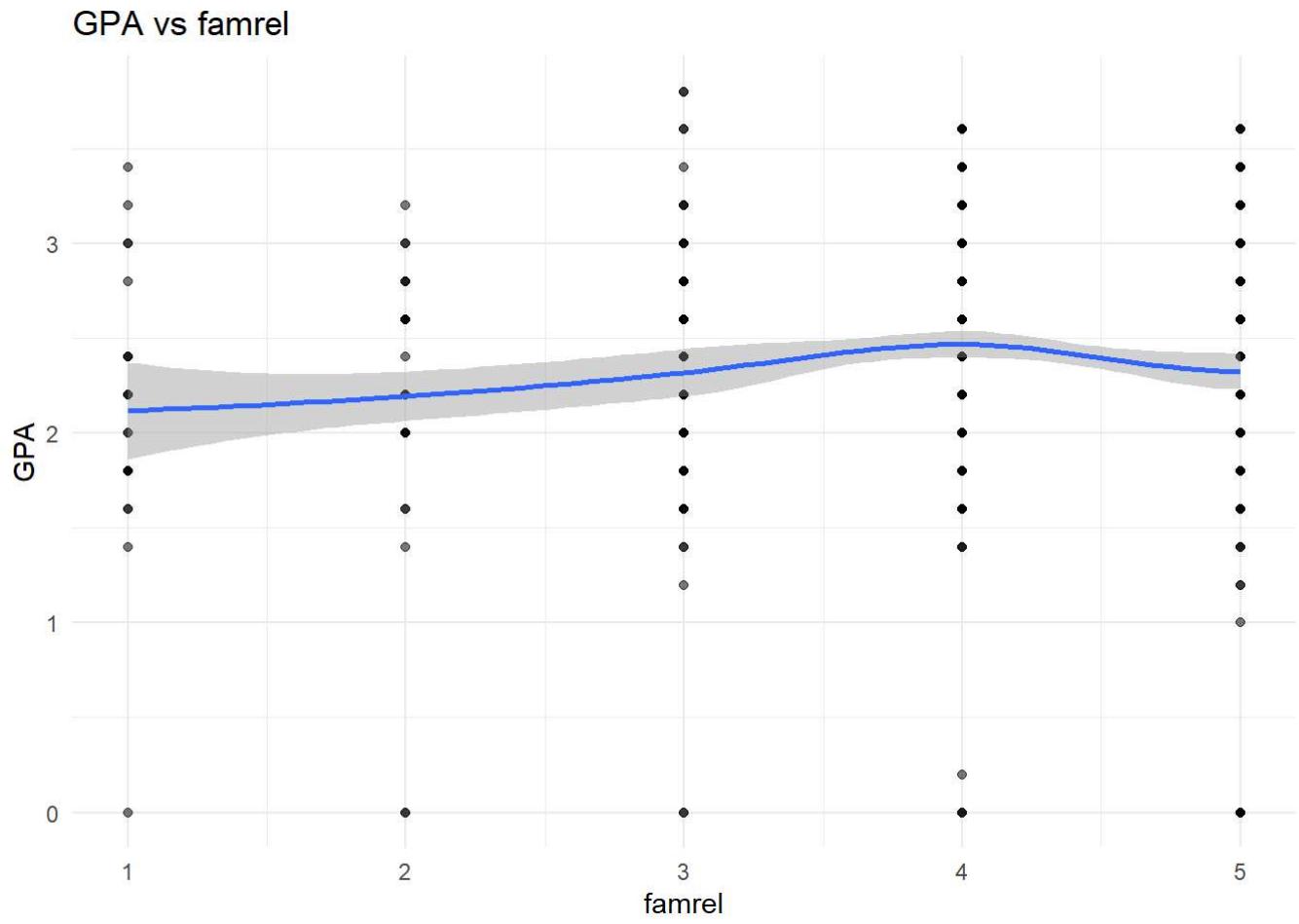

```
## `geom_smooth()` using formula = 'y ~ x'
```



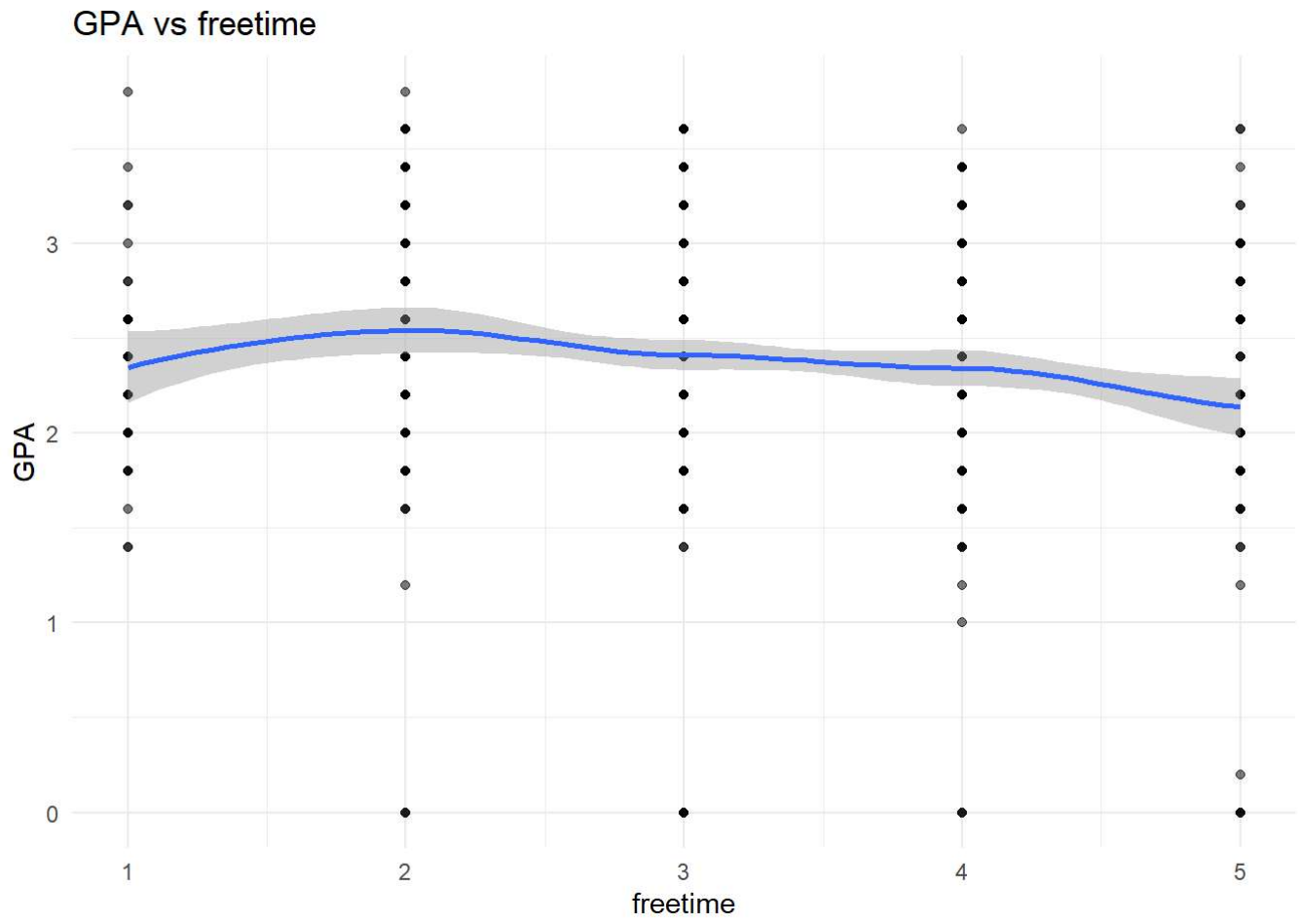

```
## `geom_smooth()` using formula = 'y ~ x'
```



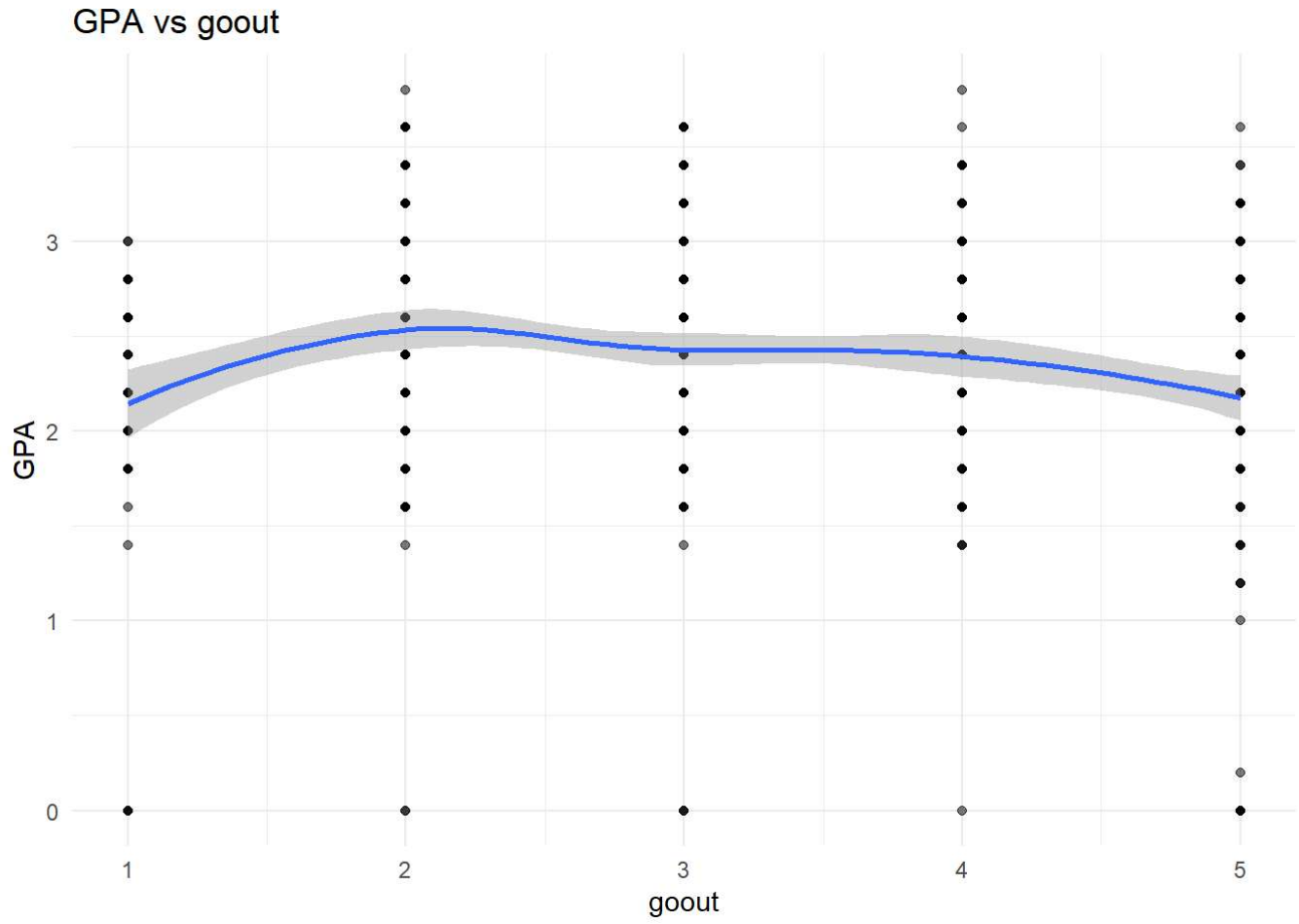

```
## `geom_smooth()` using formula = 'y ~ x'
```



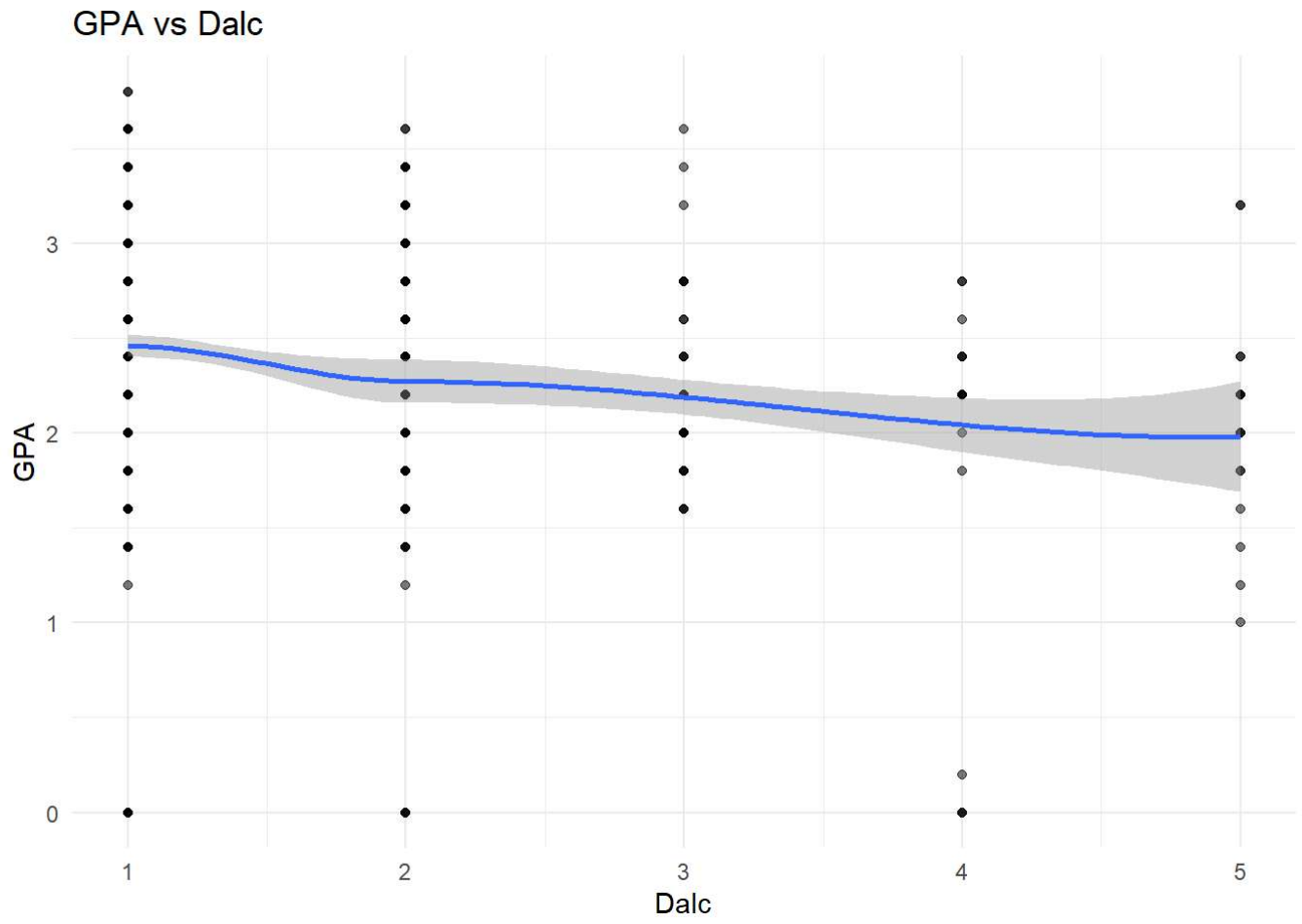

```
## `geom_smooth()` using formula = 'y ~ x'
```



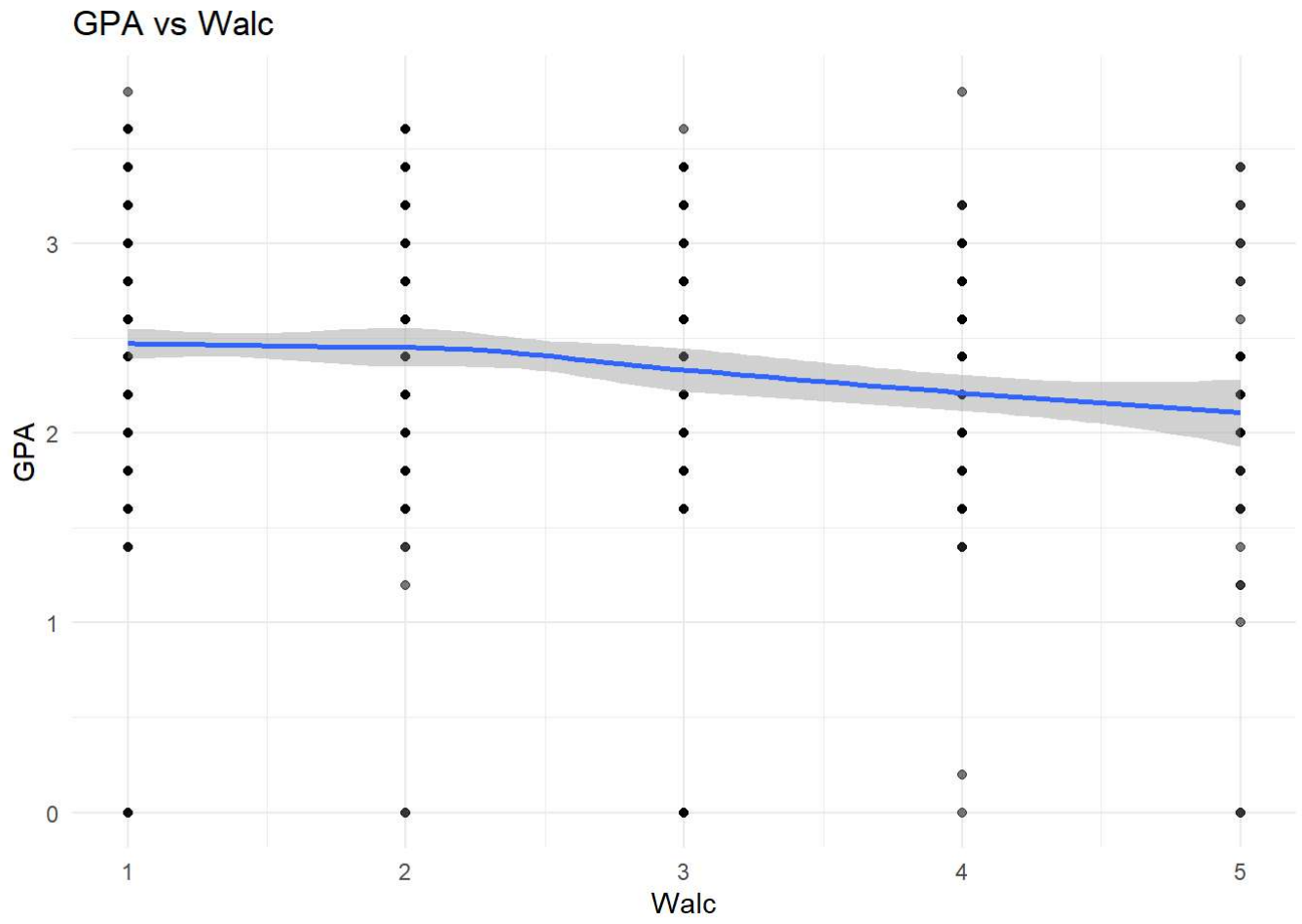

```
## `geom_smooth()` using formula = 'y ~ x'
```



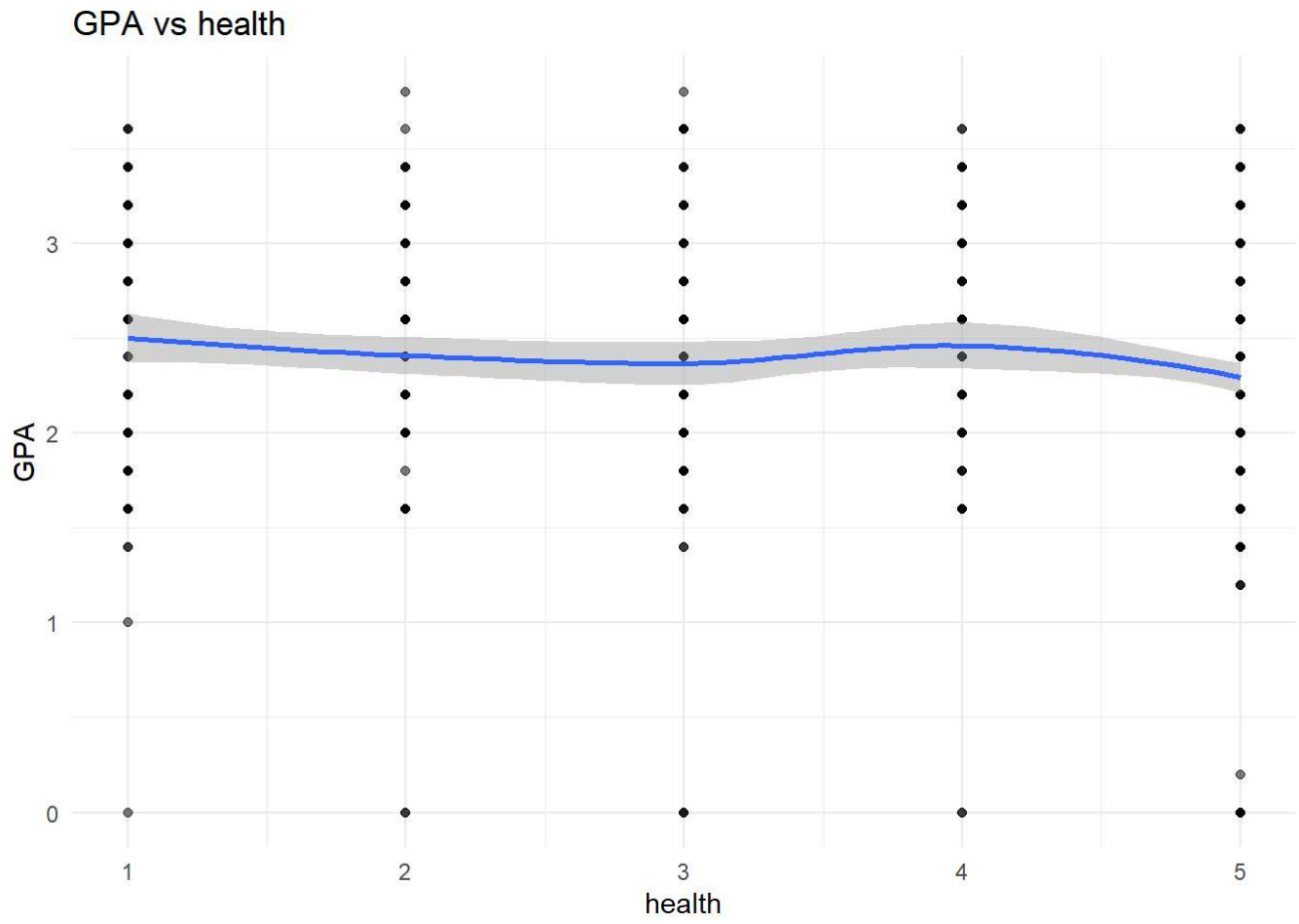

```
## `geom_smooth()` using formula = 'y ~ x'
```



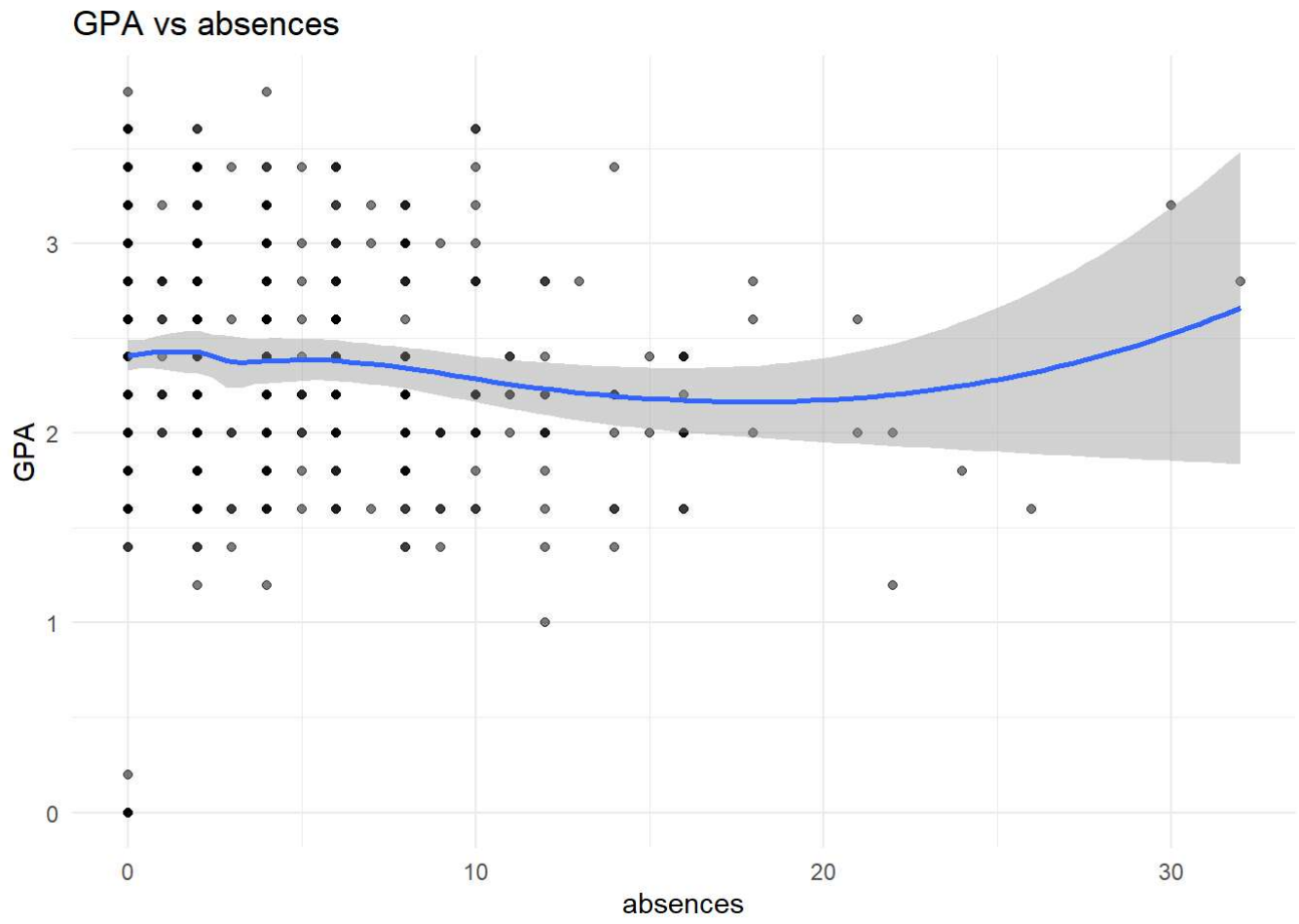

```
## `geom_smooth()` using formula = 'y ~ x'
```



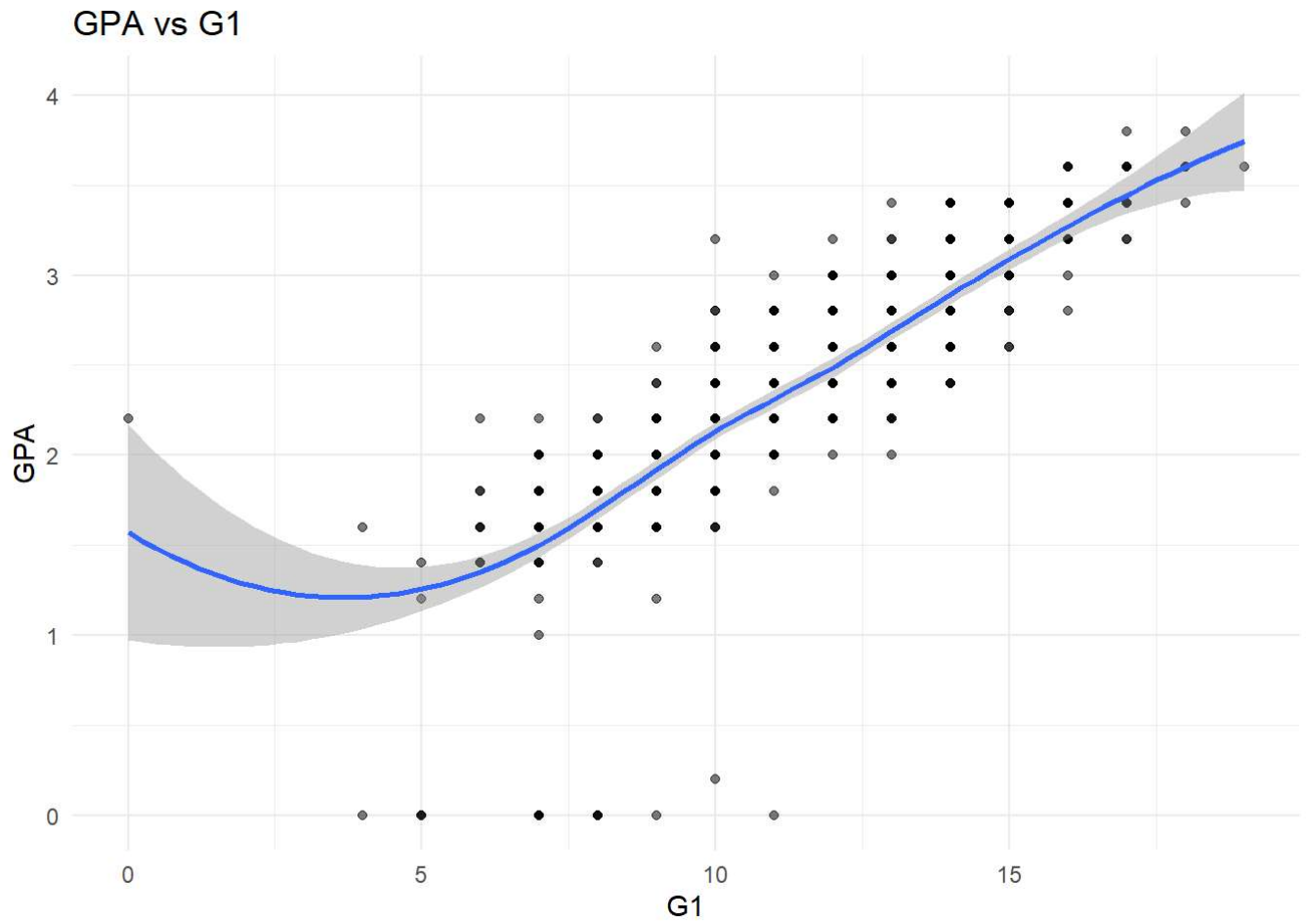

```
## `geom_smooth()` using formula = 'y ~ x'
```



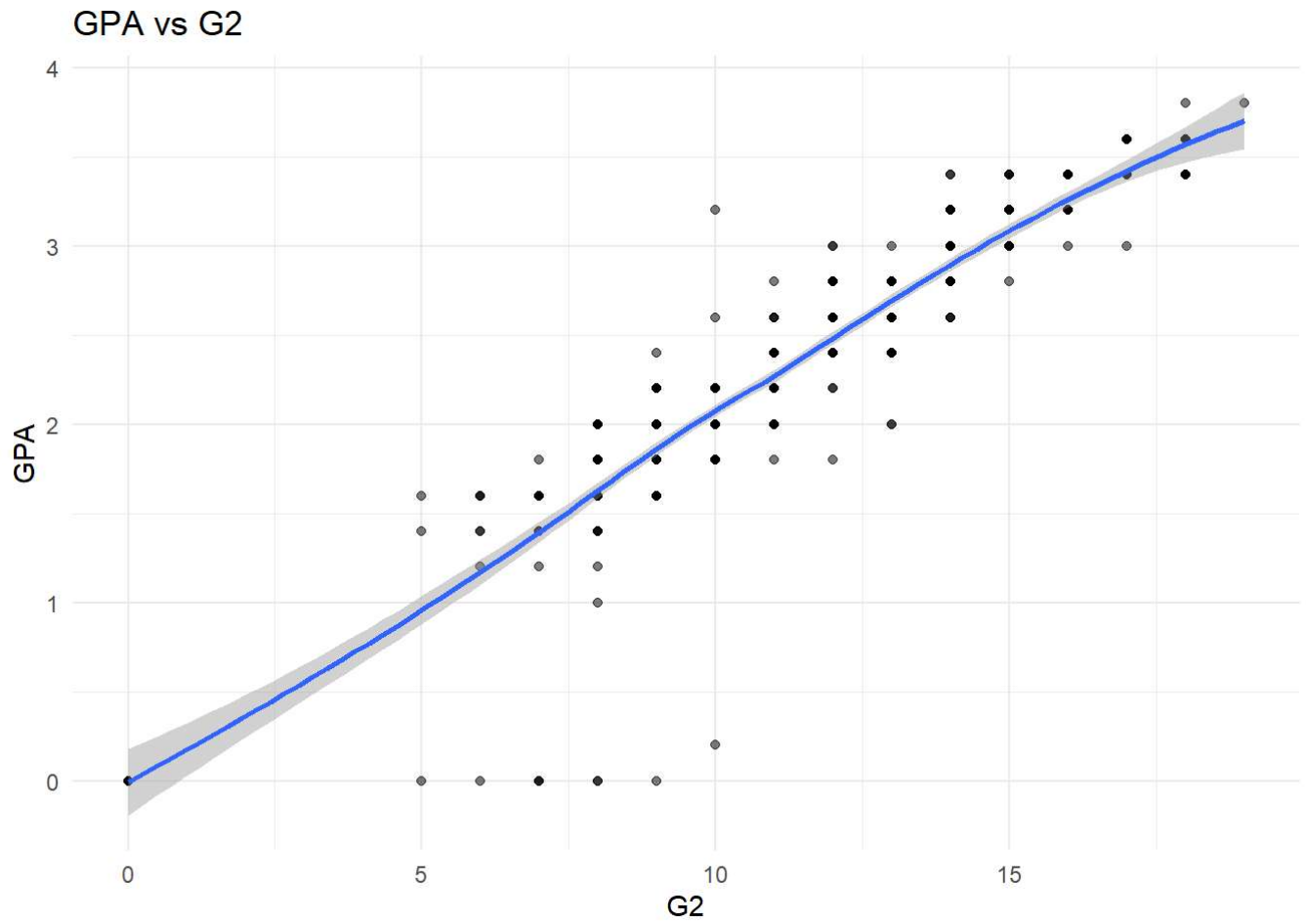

```
## `geom_smooth()` using formula = 'y ~ x'
```



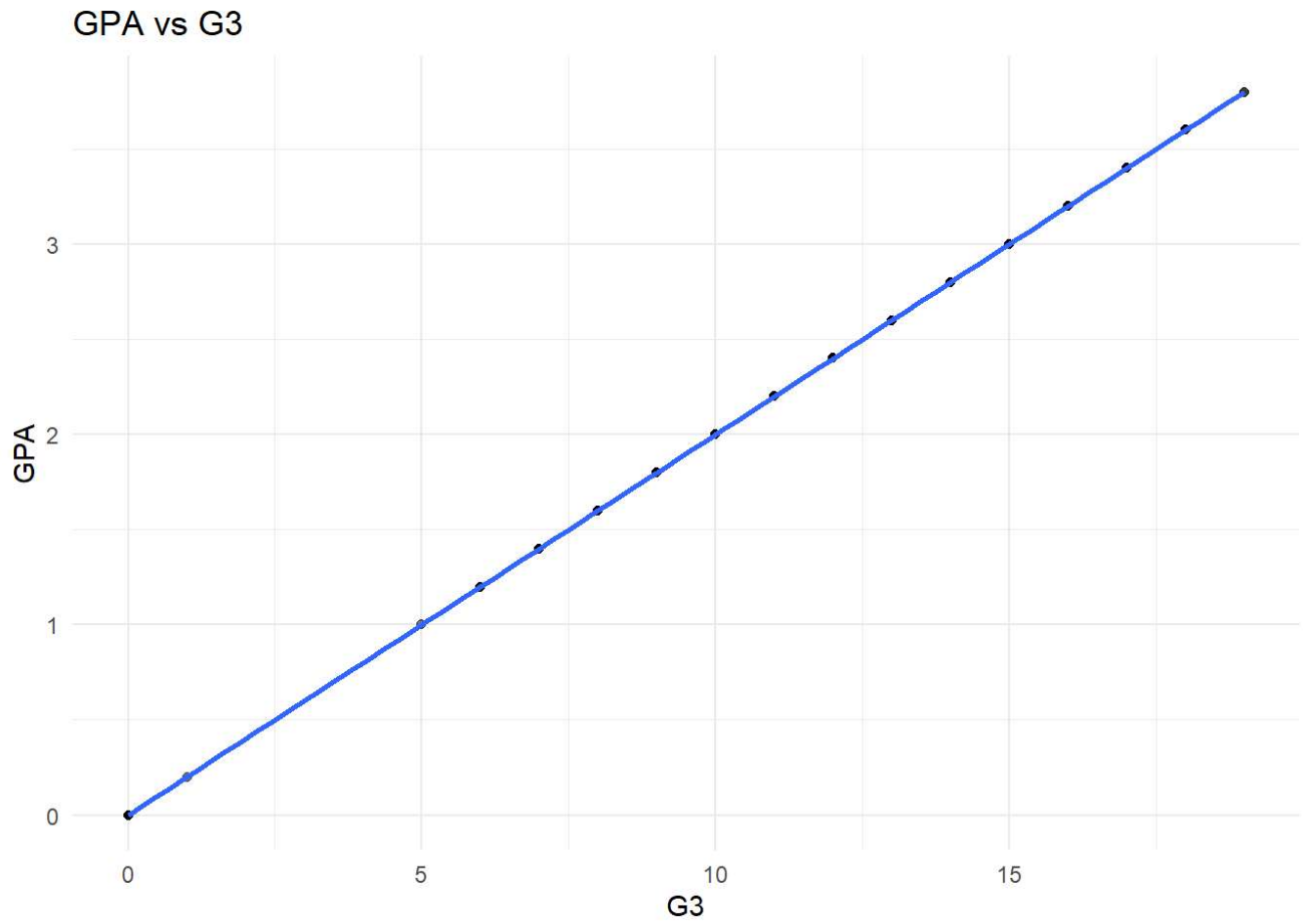

```
## `geom_smooth()` using formula = 'y ~ x'
```

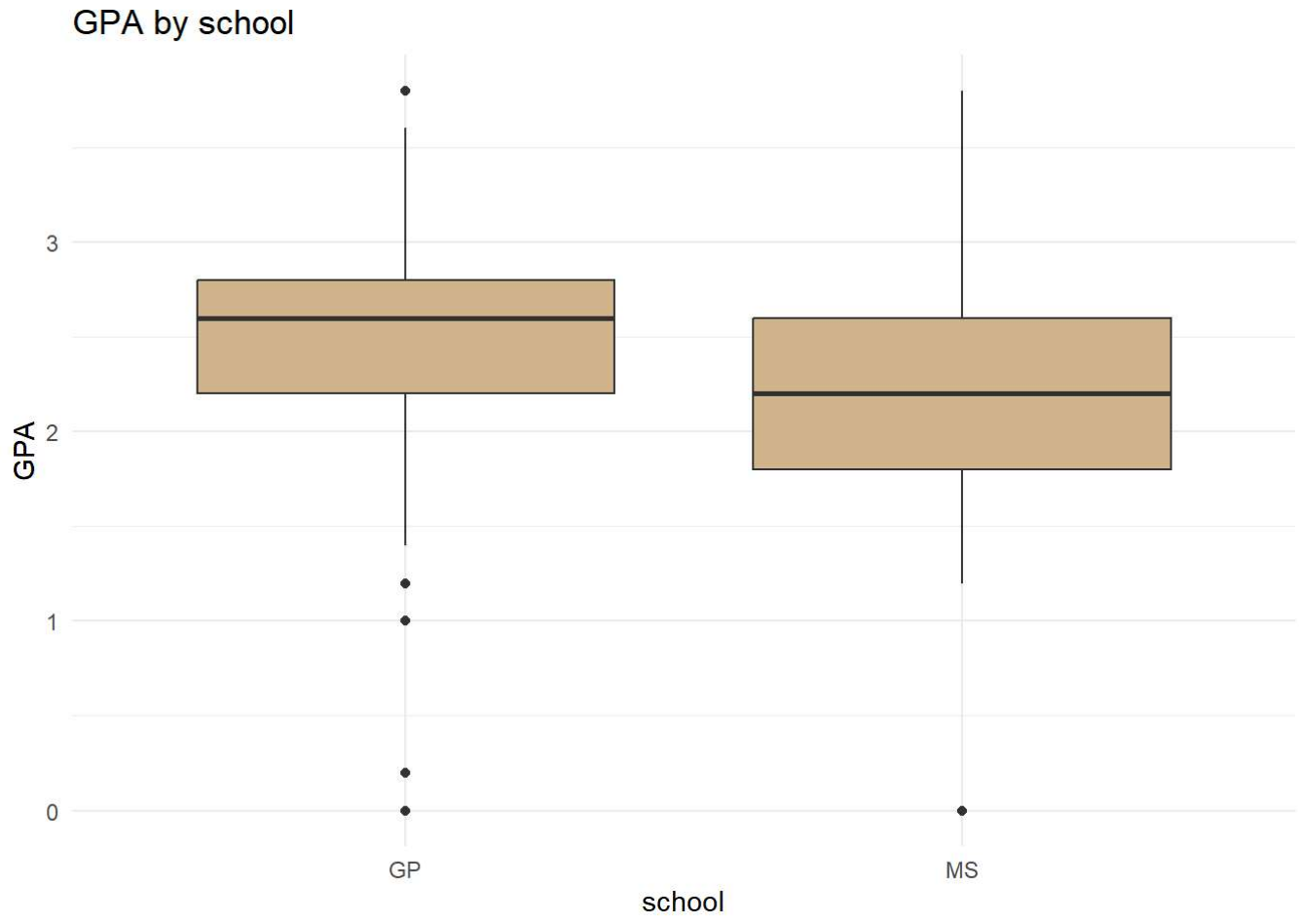


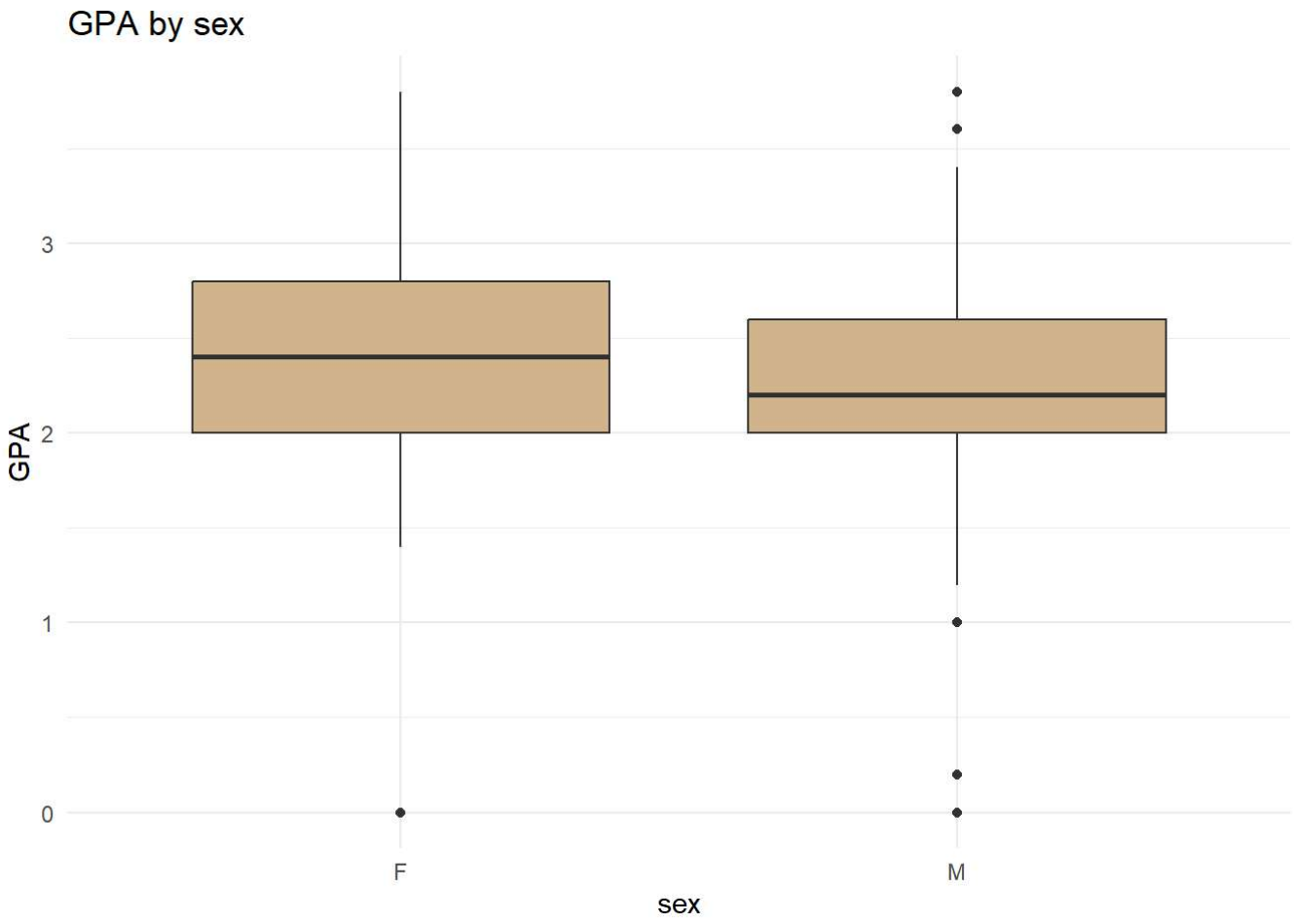

```
## `geom_smooth()` using formula = 'y ~ x'
```

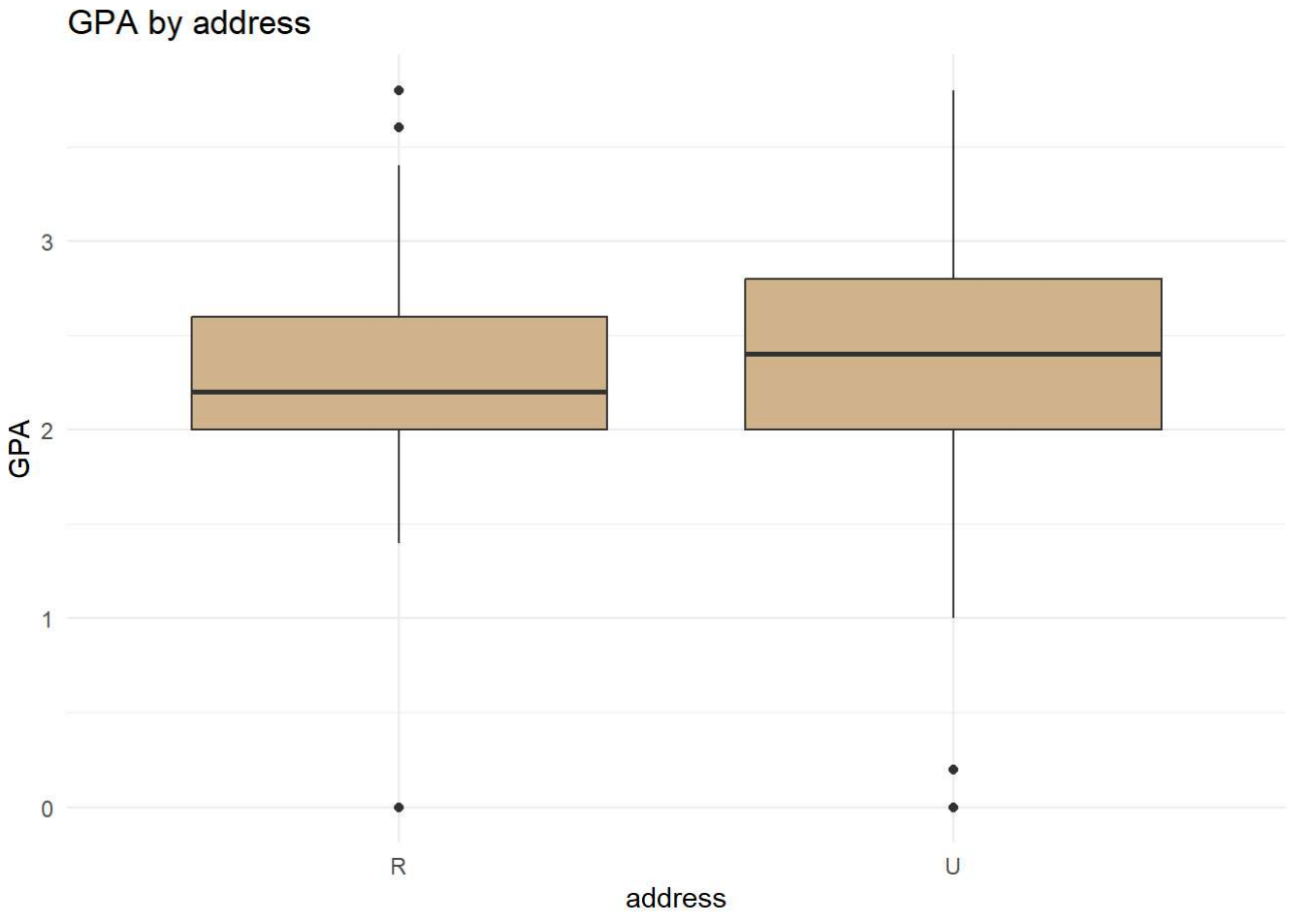


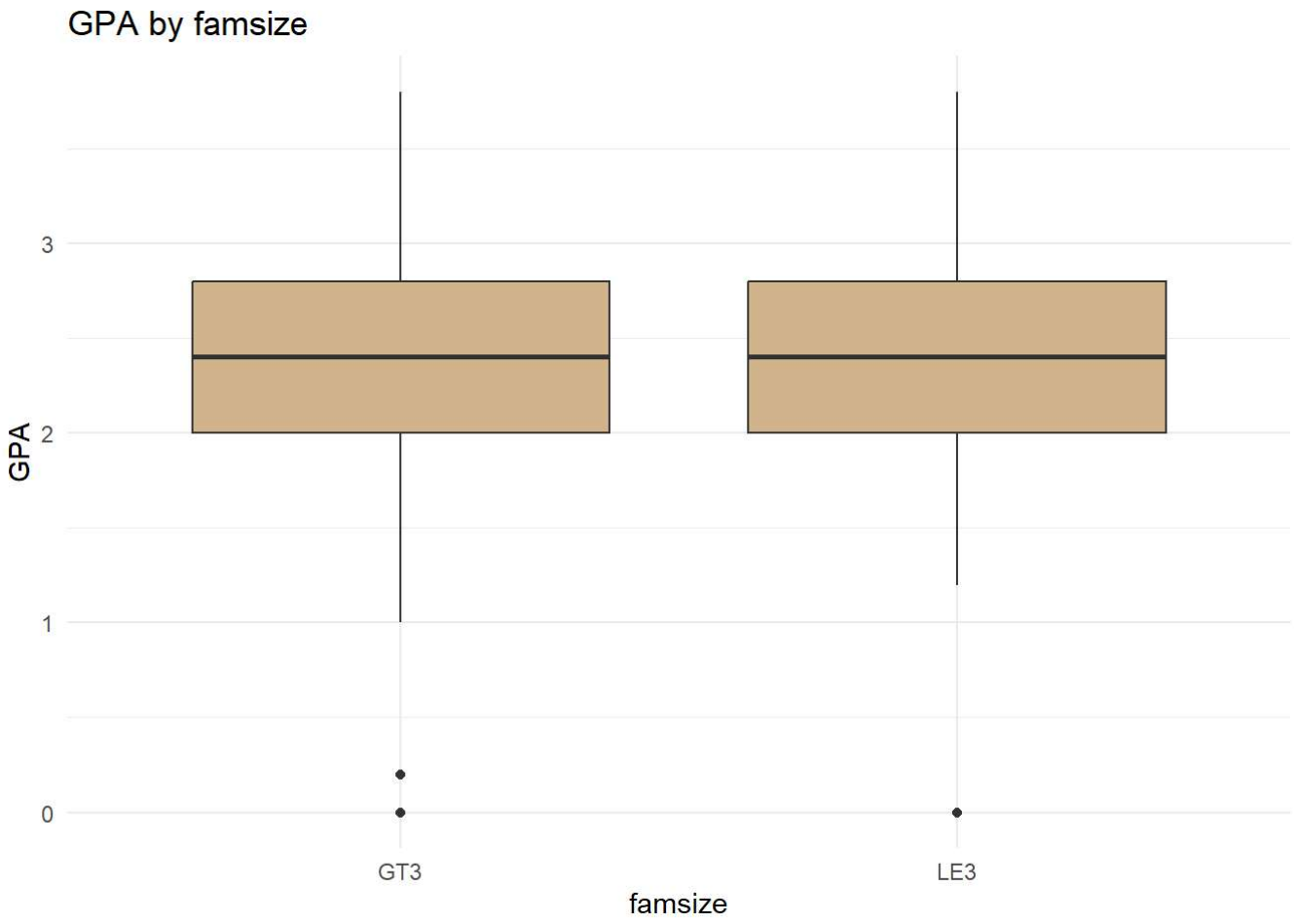

```
## `geom_smooth()` using formula = 'y ~ x'
```

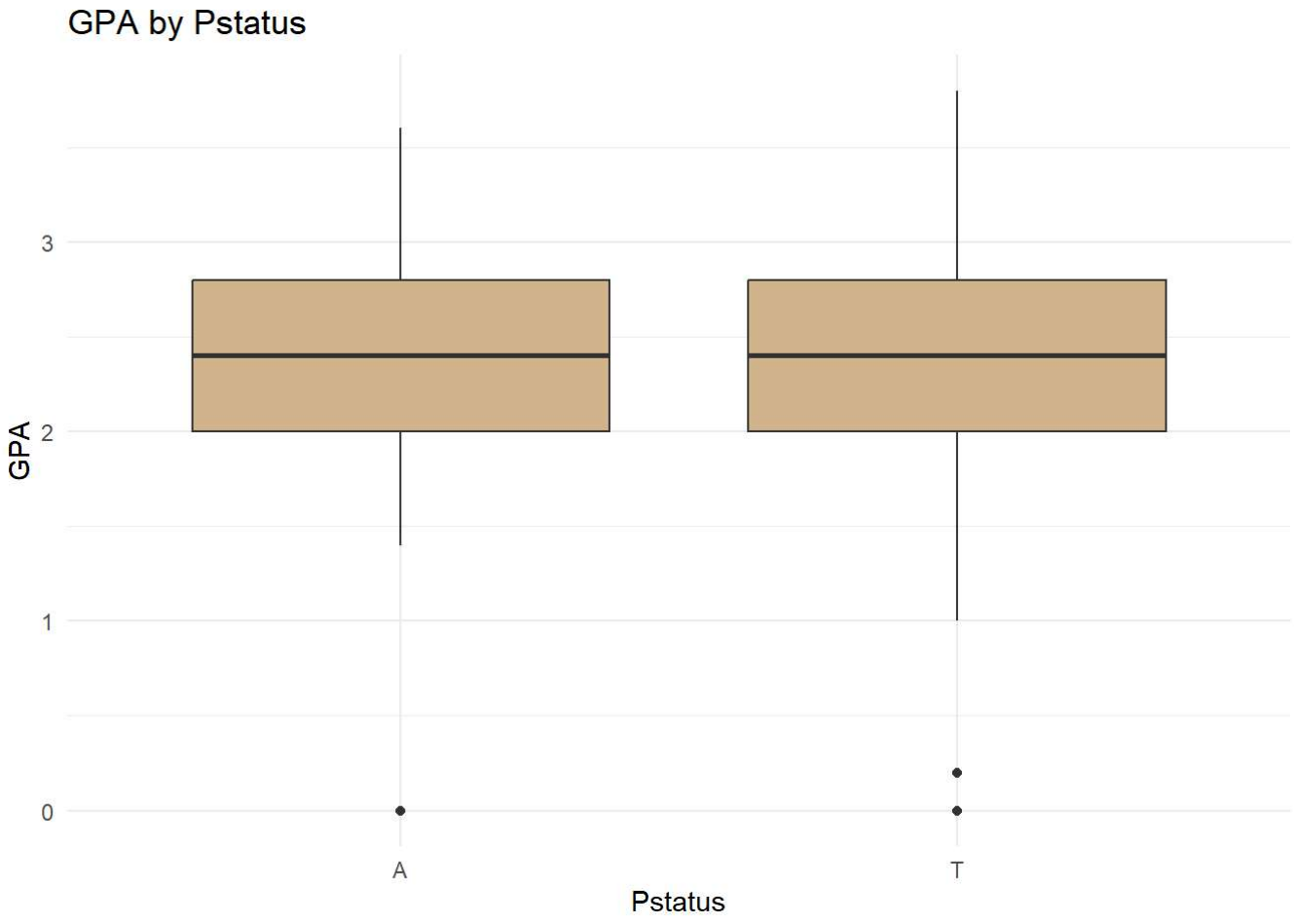

GPA vs G3
GPA
G3
0
1
2
3
0
5
10
15

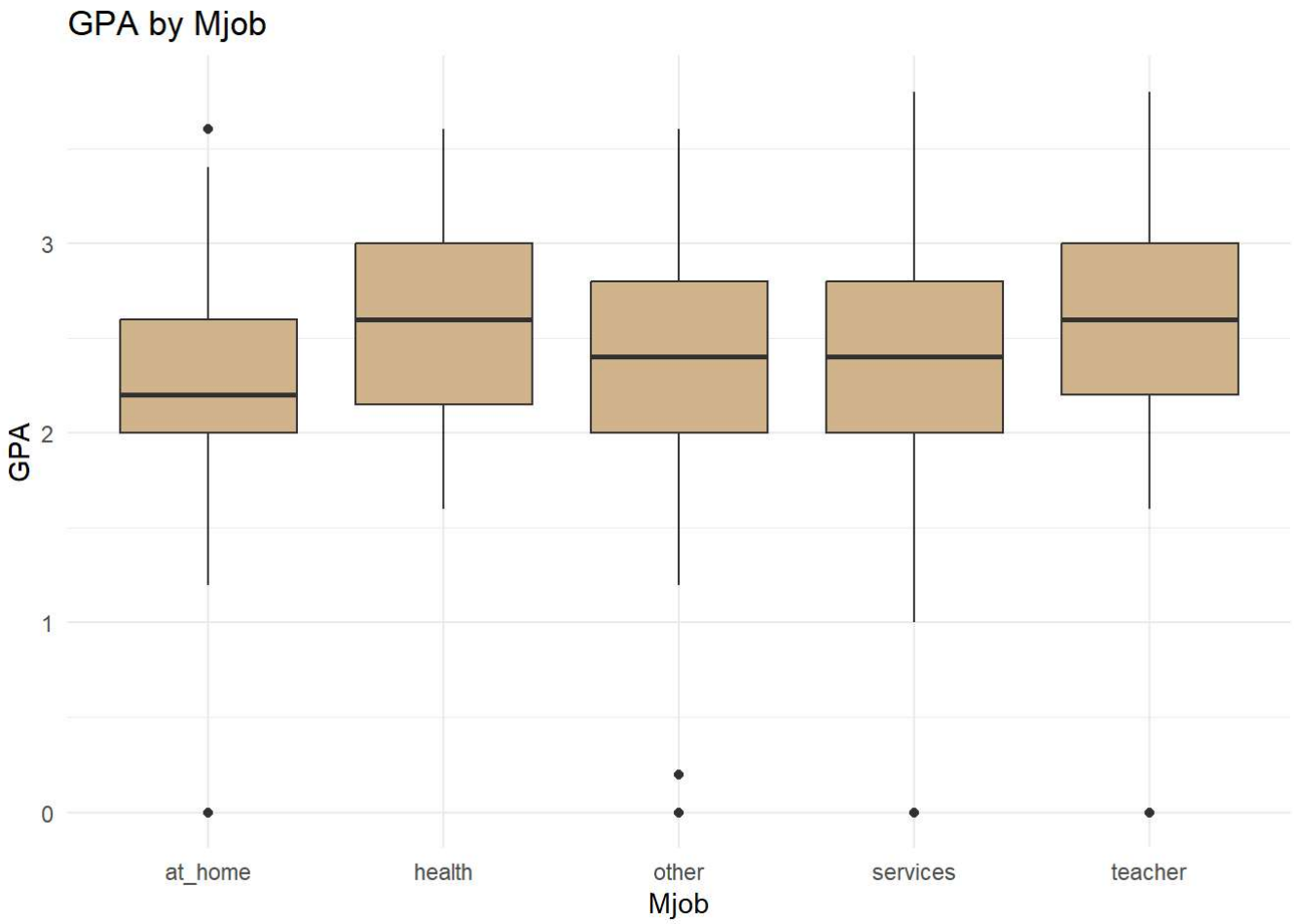

GPA by school
GPA
school
0
1
2
3
GP
MS

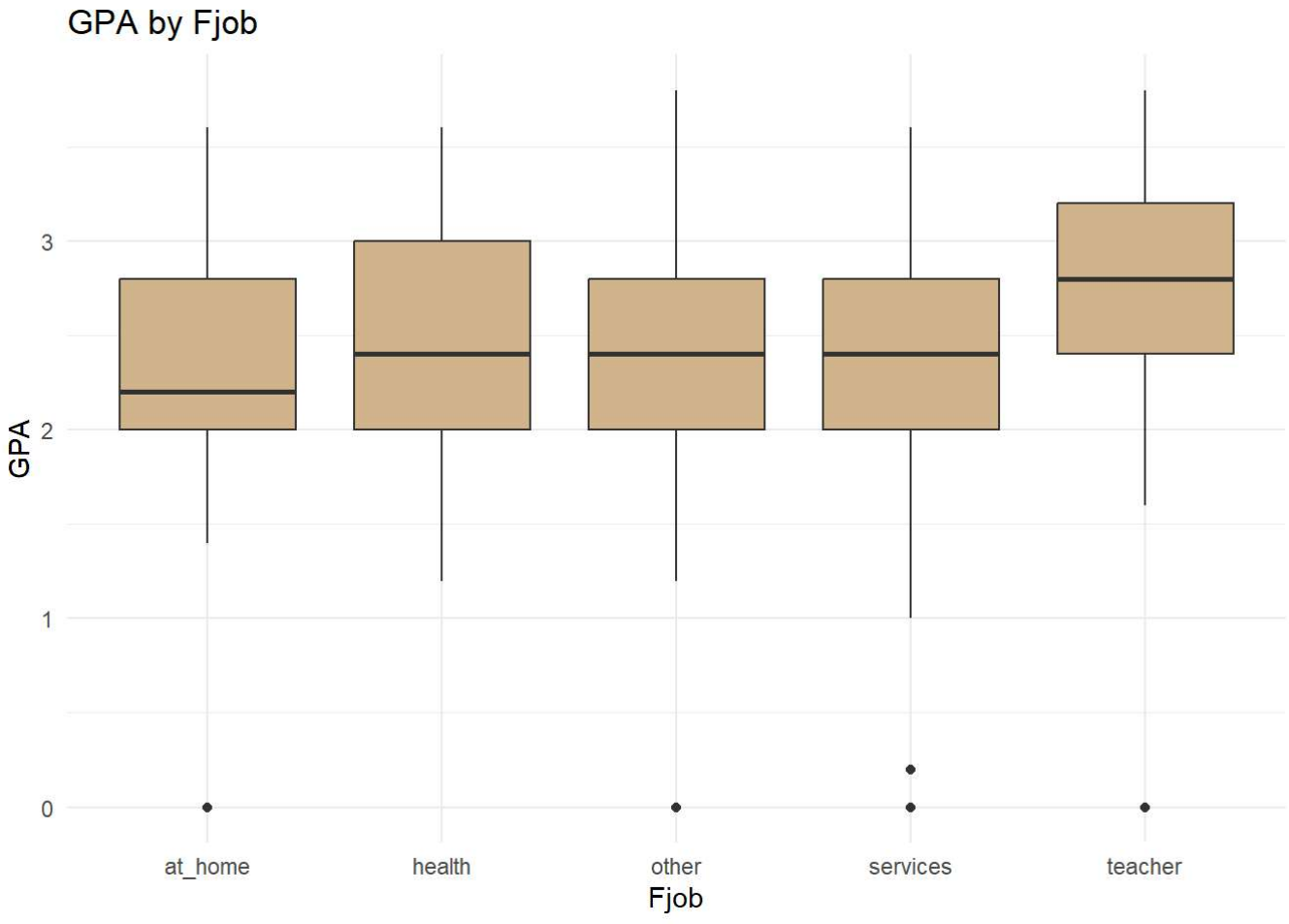

GPA by sex
GPA
3
2
1
0
F
M
sex

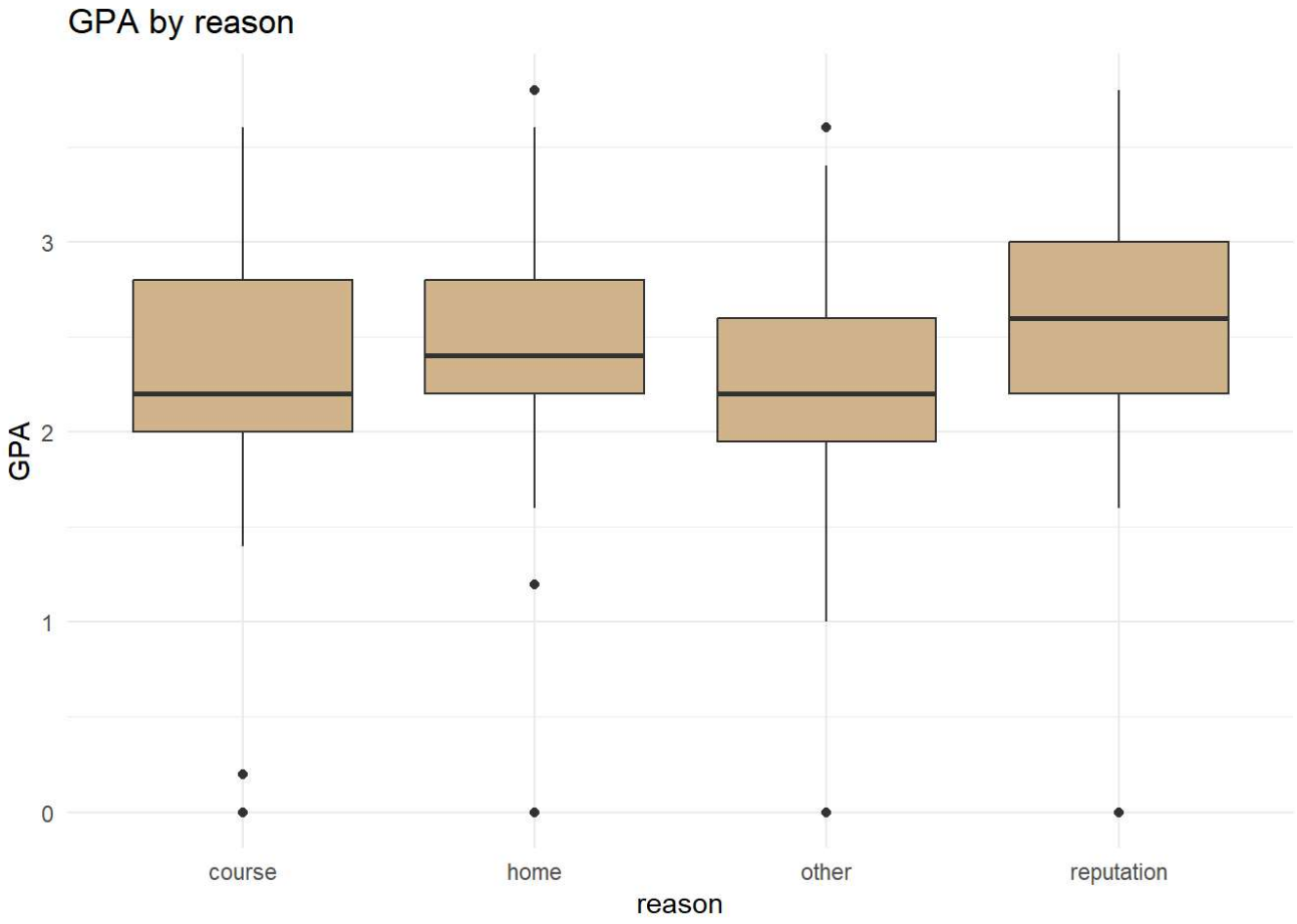

GPA by address
GPA
3
2
1
0
R
U
address

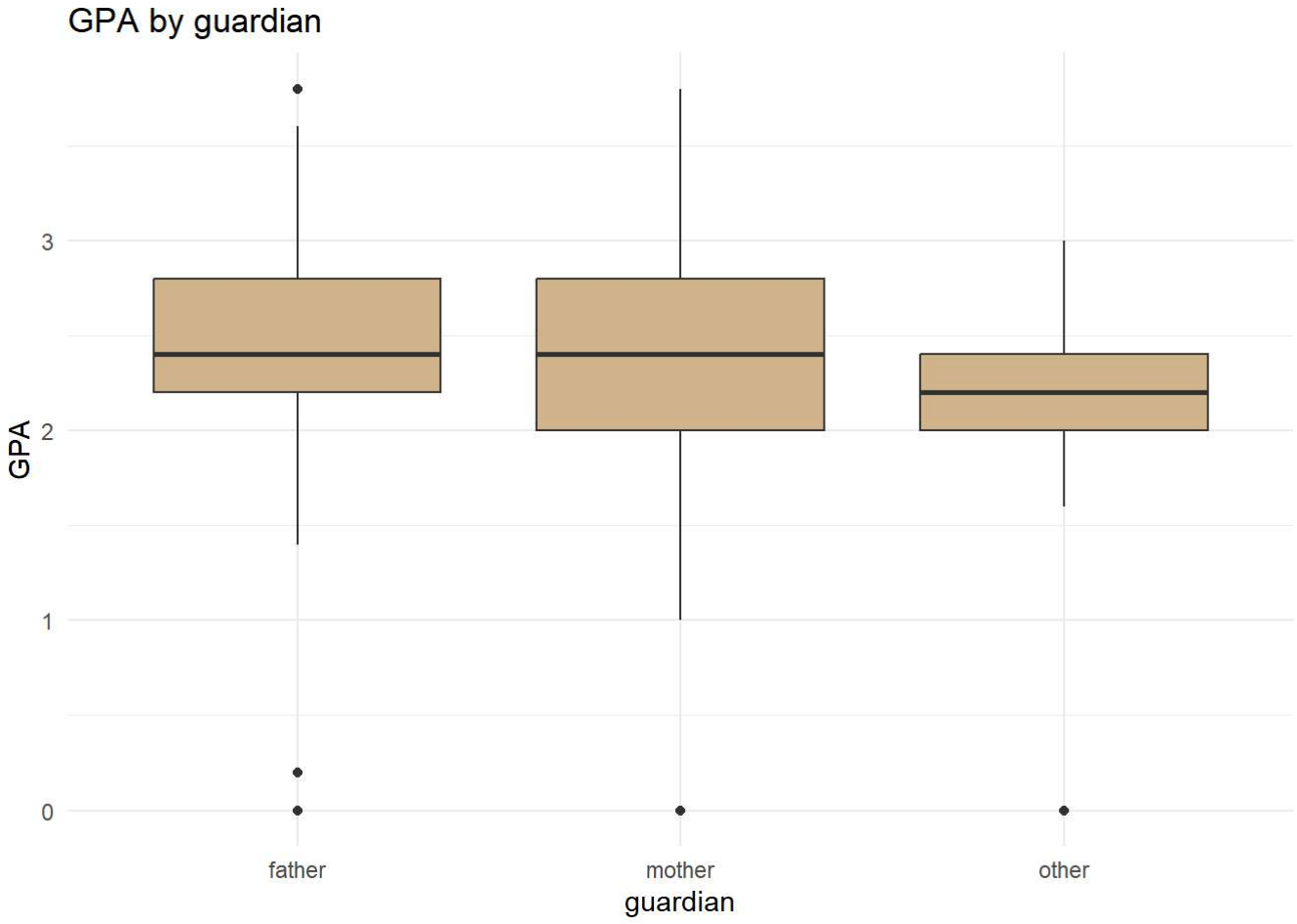

GPA by famsize
GPA
3
2
1
0
GT3
LE3
famsize

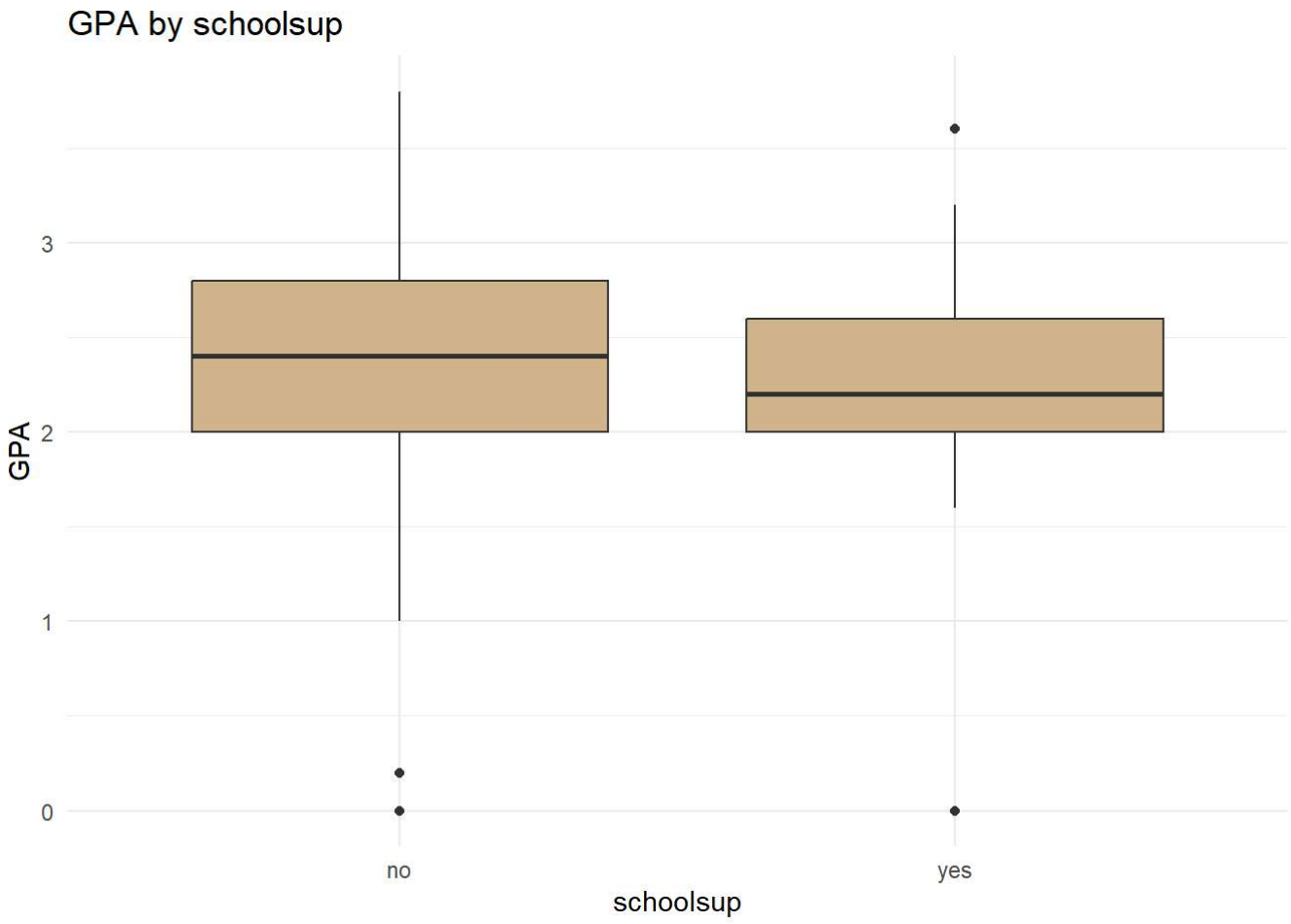

GPA by Pstatus
GPA
3
2
1
0
A
T
Pstatus

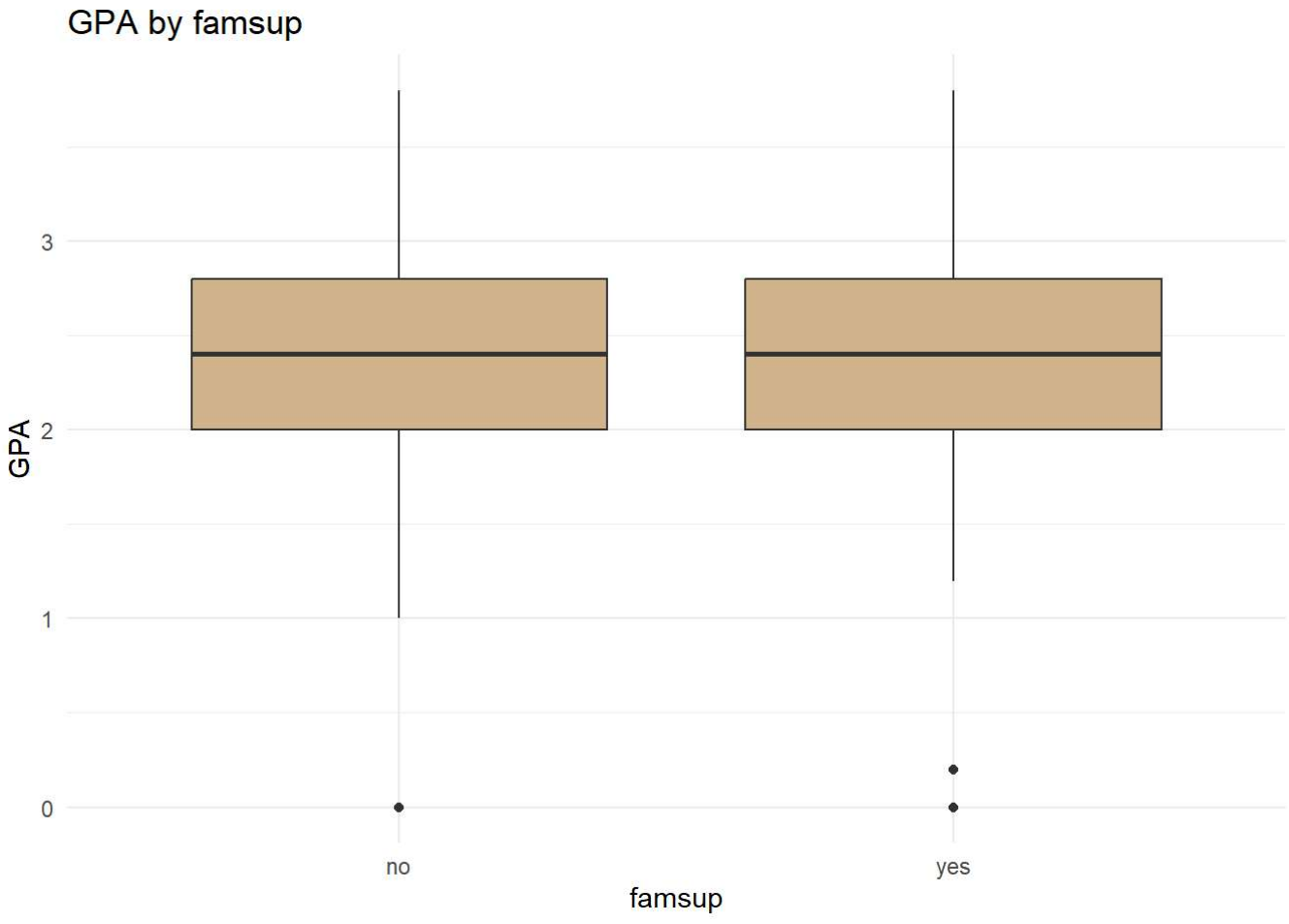

GPA by Mjob
GPA
3
2
1
0
at_home
health
other
services
teacher
Mjob

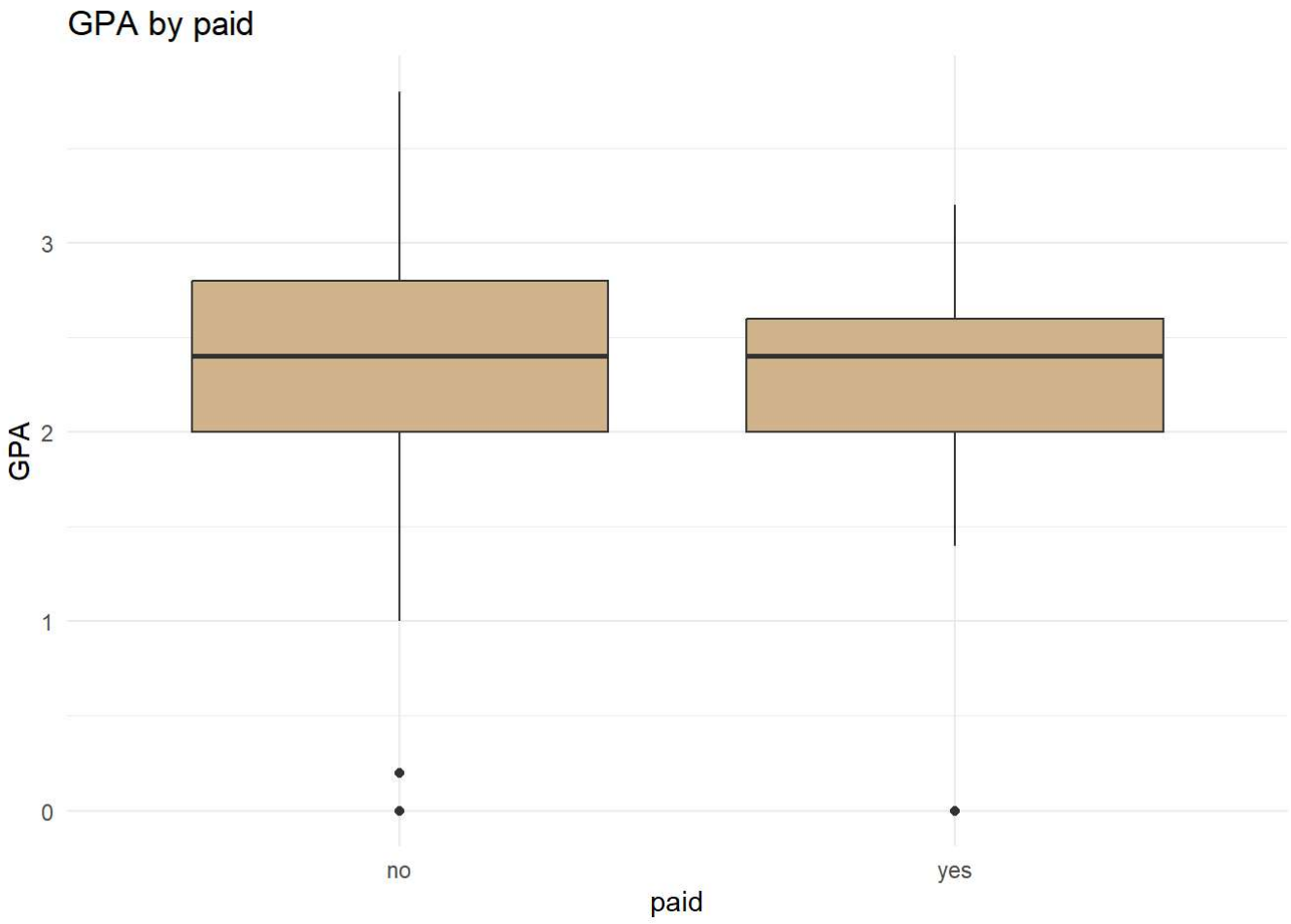

GPA by Fjob
GPA
3
2
1
0
at_home
health
other
services
teacher
Fjob

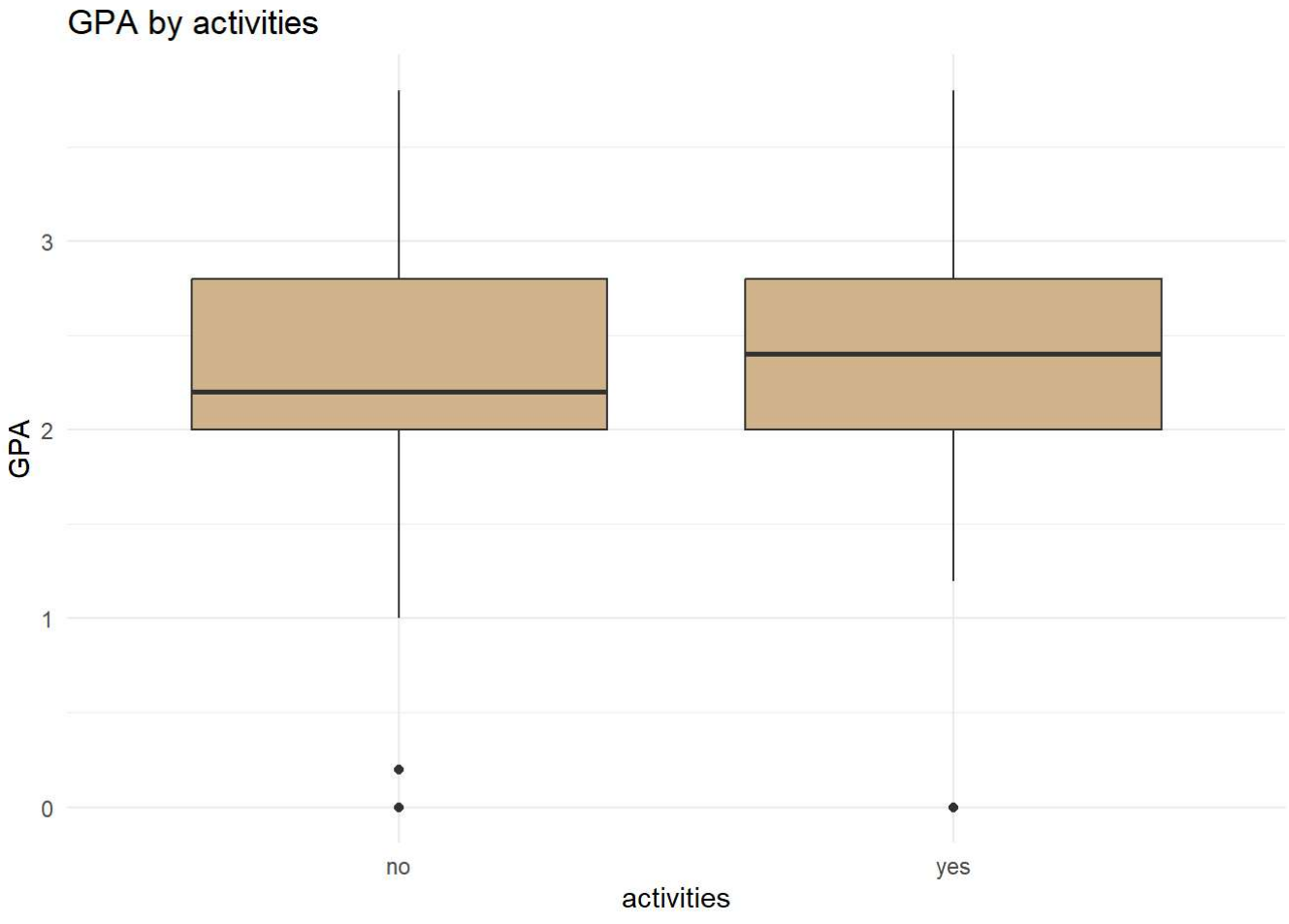

GPA by reason
GPA
3
2
1
0
course
home
other
reputation
reason

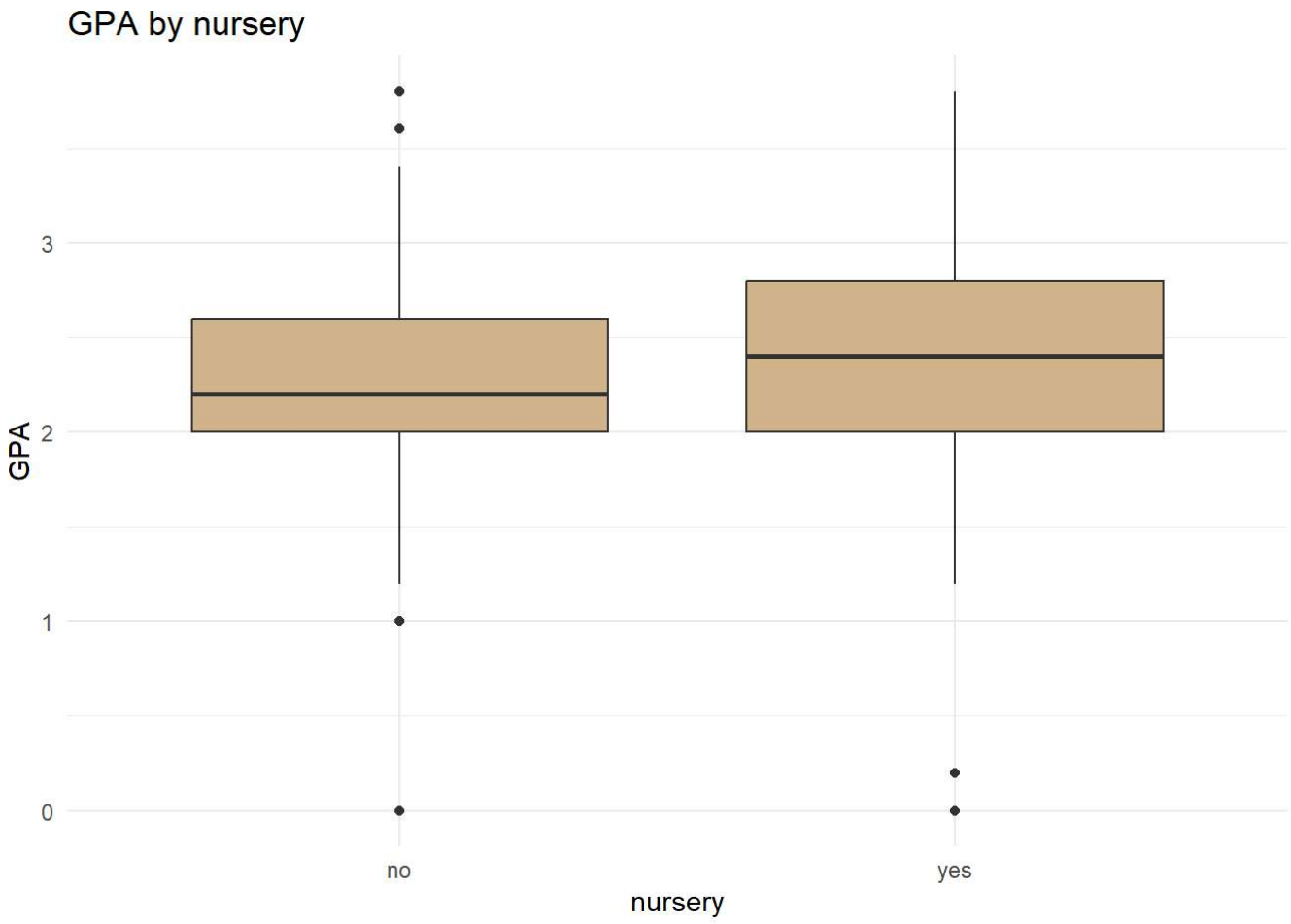

GPA by guardian
GPA
3
2
1
0
father
mother
other
guardian

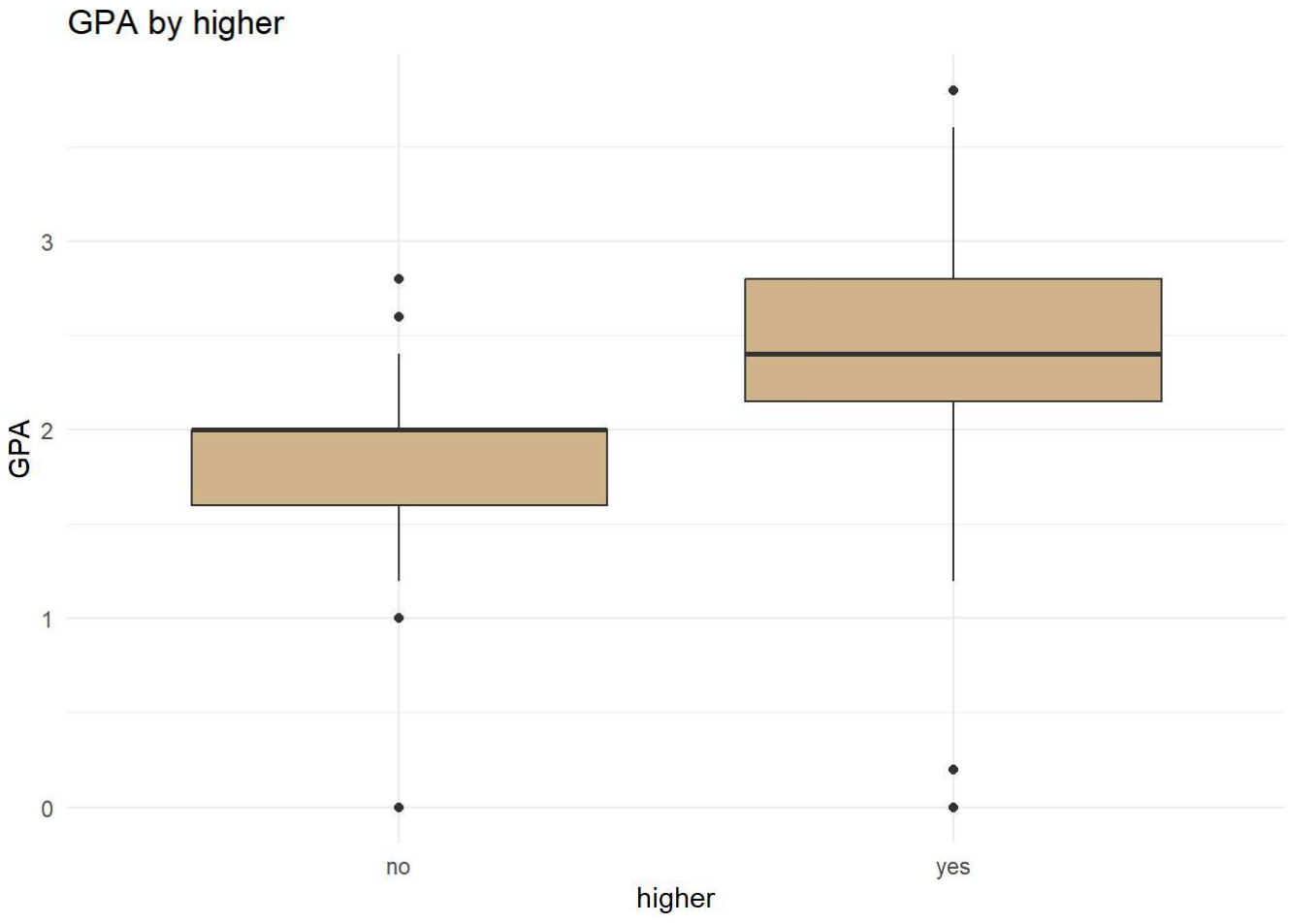

GPA by schoolsup
GPA
3
2
1
0
no
yes
schoolsup

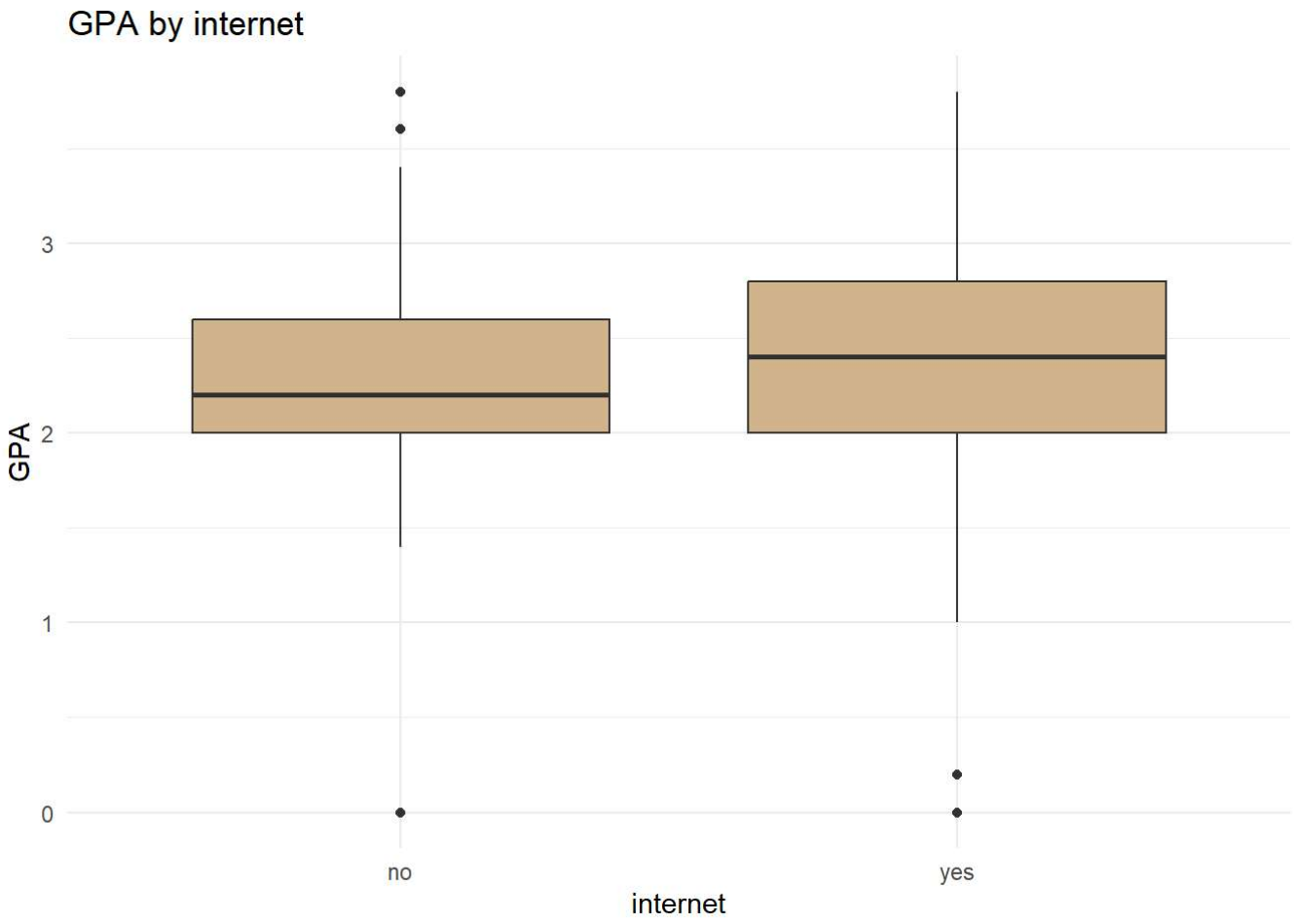

GPA by famsup
GPA
3
2
1
0
no
yes
famsup

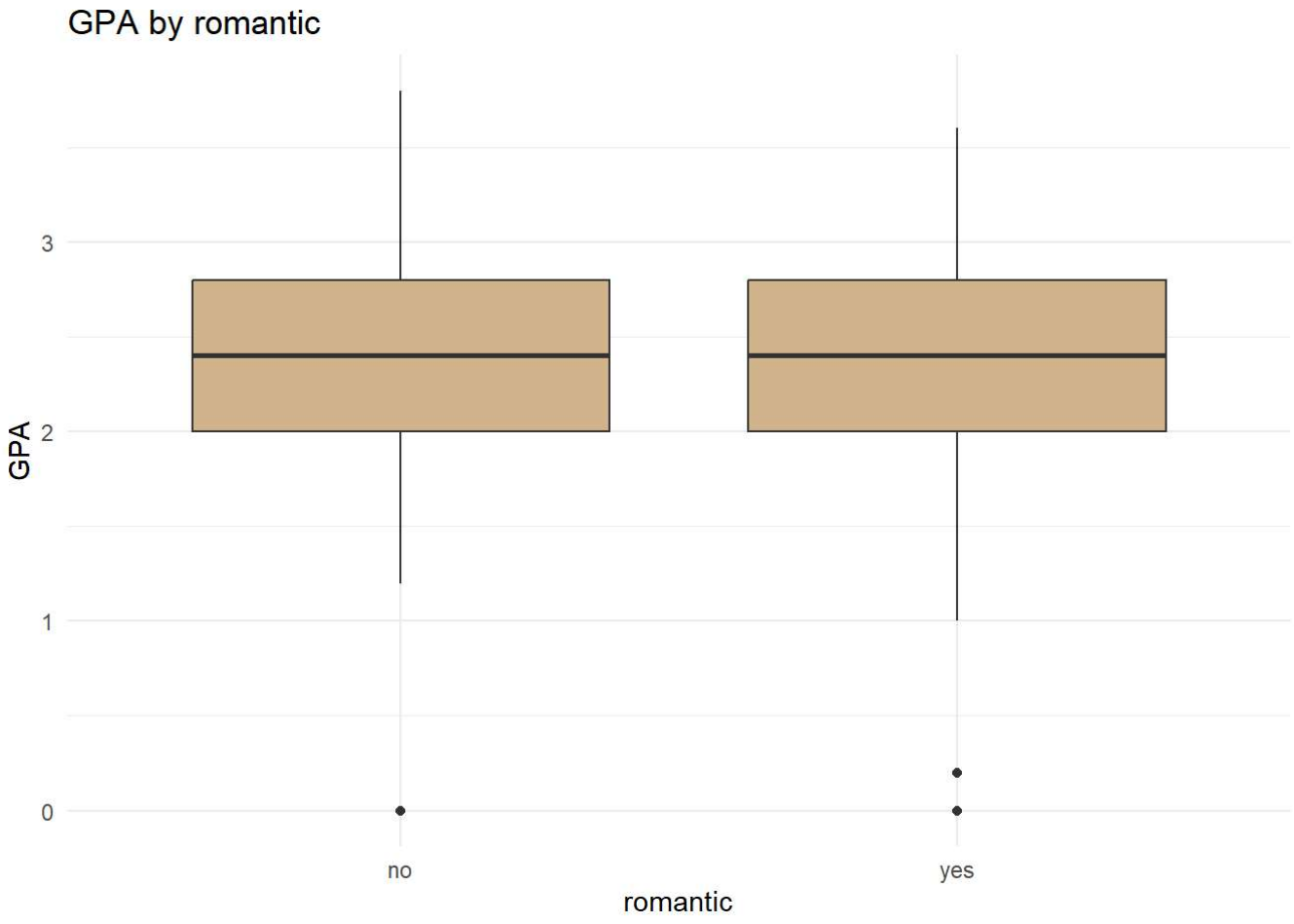

GPA by paid
GPA
3
2
1
0
no
yes
paid


GPA by activities
GPA
3
2
1
0
no
yes
activities


GPA by nursery
GPA
3
2
1
0
no
yes
nursery


GPA by higher
GPA
3
2
1
0
no
yes
higher


GPA by internet
GPA
3
2
1
0
no
yes
internet


GPA by romantic
GPA
3
2
1
0
no
yes
romantic

```
##
##  GP  MS
## 423 226
##
##   F   M
## 383 266
##
##   R   U
## 197 452
##
## GT3 LE3
## 457 192
##
##   A   T
##  80 569
##
##  at_home   health    other services  teacher
##      135       48      258      136       72
##
##  at_home   health    other services  teacher
##       42       23      367      181       36
##
##     course       home      other reputation
##        285        149         72        143
##
## father mother  other
##    153    455     41
##
##  no yes
## 581  68
##
##  no yes
## 251 398
##
##  no yes
## 610  39
##
##  no yes
## 334 315
##
##  no yes
## 128 521
##
##  no yes
##  69 580
##
##  no yes
## 151 498
##
##  no yes
## 410 239
```

```
# Checking for outliers in absences variable
Q1 <- quantile(port_students$absences, 0.25)
Q3 <- quantile(port_students$absences, 0.75)
IQR_value <- IQR(port_students$absences)
lower_bound <- Q1 - 1.5 * IQR_value # returns -9, but this can be represented as 0.
upper_bound <- Q3 + 1.5 * IQR_value # 15
which(port_students$absences < lower_bound | port_students$absences > upper_bound) # returns ind
ices
```

```
##  [1]  41 104 151 156 162 198 207 212 213 218 231 254 255 257 264 312 326 327 398
## [20] 406 414
```

```
print("These are the extreme outliers in absences: ")
```

```
## [1] "These are the extreme outliers in absences: "
```

```
port_students$absences[which(port_students$absences < lower_bound | port_students$absences > upp
er_bound)]
```

```
##  [1] 16 16 24 22 16 32 16 16 30 21 16 18 16 26 16 16 22 18 18 16 21
```

```
print("We may need to perform log transformation. We'll evaluate its quantile regression results
before any transformation.")
```

```
## [1] "We may need to perform log transformation. We'll evaluate its quantile regression result
s before any transformation."
```

```
# 21 extreme outliers
```

Pairwise relationships

```
# This can be slow on large datasets
#GGally::ggpairs(port_students[, c("GPA", "studytime", "failures", "absences")])
```

```
# Write data to csv file
port_students <- port_students %>% dplyr::select(-G3) # remove G3
write.csv(port_students, "port_students_revised.csv", row.names = FALSE)
```