script

2025-12-04

Exploratory data analysis

```
# Load packages
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(ggplot2)
```

Reading in the dataset:

```
dat <- read.csv("student-mat.csv", sep = ";")
str(dat)
```

```
## 'data.frame':    395 obs. of  33 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 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ 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  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Organizing into relevant cols:

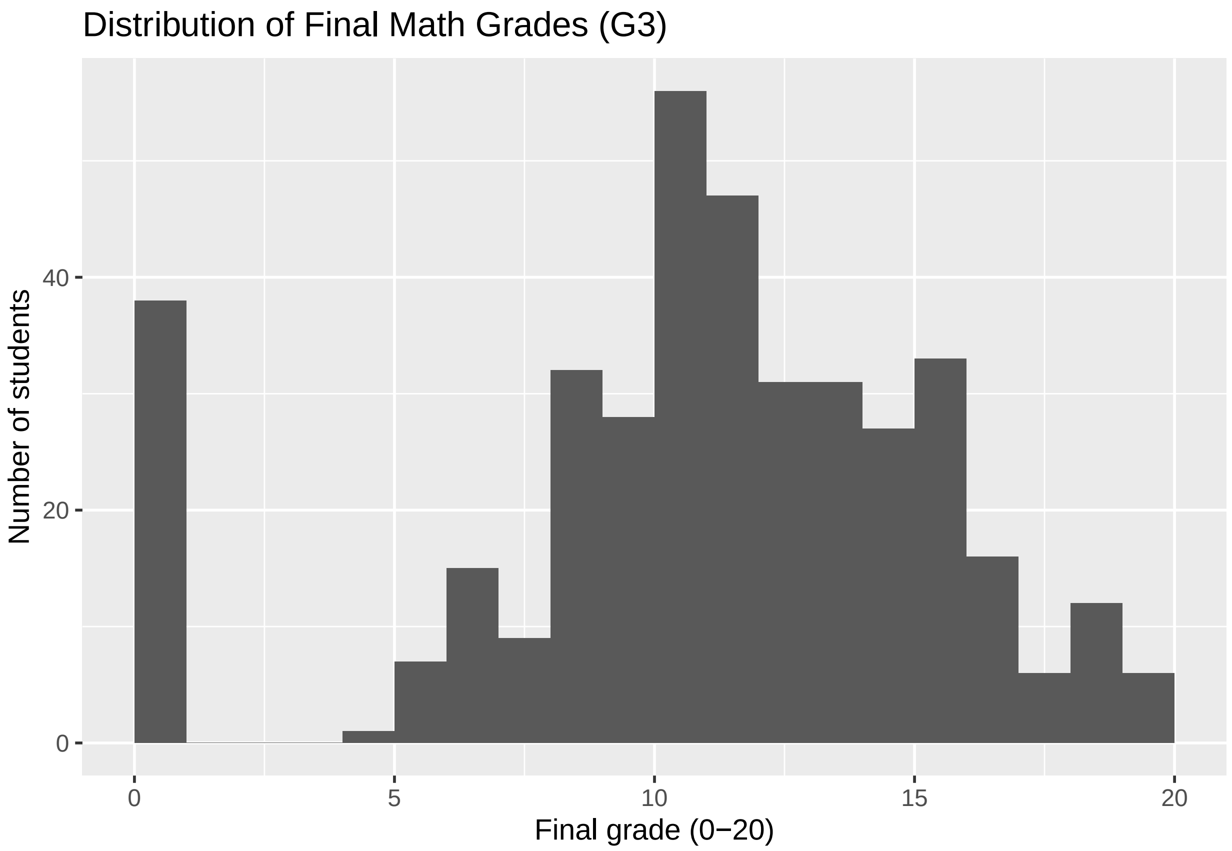
```r
dat <- dat %>%
  mutate(
    sex      = factor(sex),
    school   = factor(school),
    address  = factor(address),
    famsize  = factor(famsize),
    Pstatus  = factor(Pstatus),
    internet = factor(internet),
    romantic = factor(romantic),
    studytime = factor(
      studytime,
      levels = 1:4,
      labels = c("low", "moderate", "high", "very high")
    )
  )

# Basic summary of numeric variables
summary(select(dat, G1, G2, G3, age, absences))
```

```
##        G1              G2              G3             age      
##  Min.   : 3.00   Min.   : 0.00   Min.   : 0.00   Min.   :15.0  
##  1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:16.0  
##  Median :11.00   Median :11.00   Median :11.00   Median :17.0  
##  Mean   :10.91   Mean   :10.71   Mean   :10.42   Mean   :16.7  
##  3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:18.0  
##  Max.   :19.00   Max.   :19.00   Max.   :20.00   Max.   :22.0  
##     absences     
##  Min.   : 0.000  
##  1st Qu.: 0.000  
##  Median : 4.000  
##  Mean   : 5.709  
##  3rd Qu.: 8.000  
##  Max.   :75.000  
```

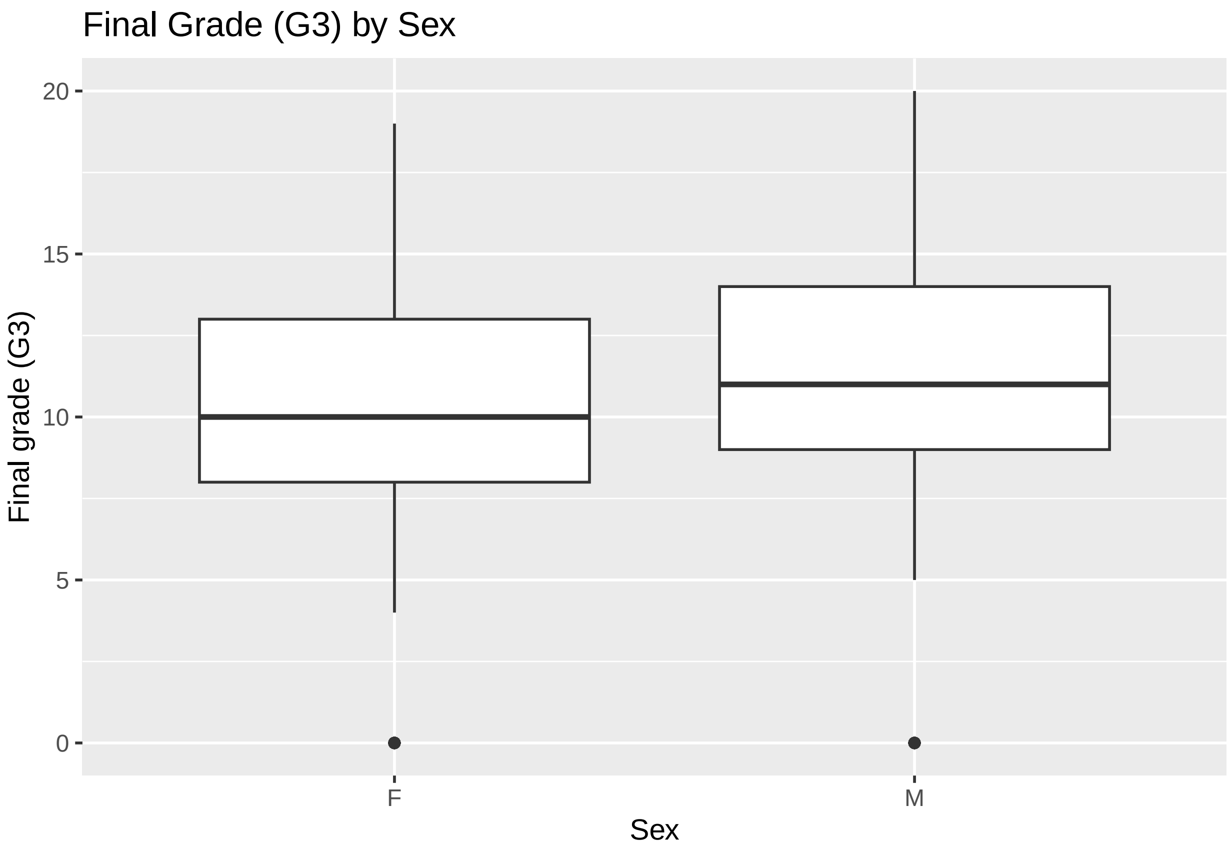
Distribution of math grades?

```r
ggplot(dat, aes(x = G3)) +
  geom_histogram(binwidth = 1, boundary = 0, closed = "left") +
  labs(
    title = "Distribution of Final Math Grades (G3)",
    x = "Final grade (0-20)",
    y = "Number of students"
  )
```



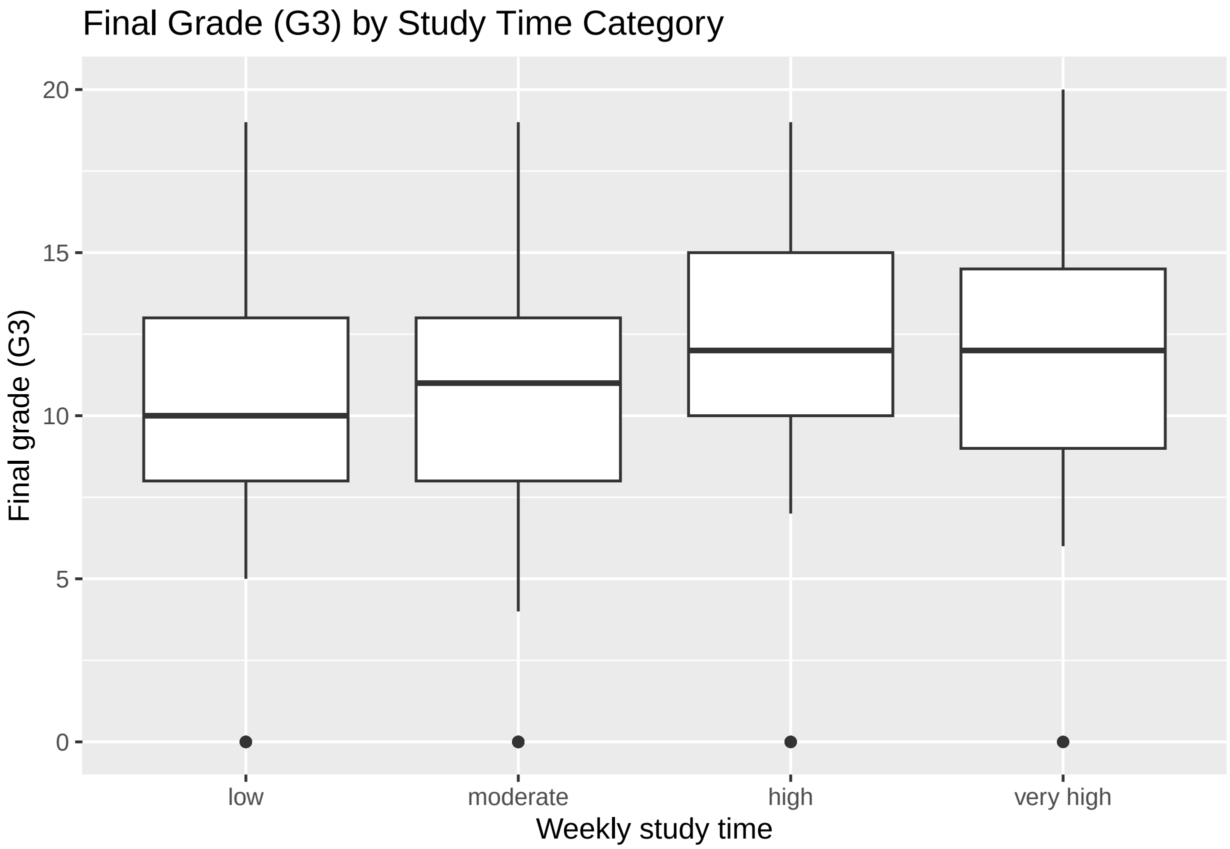
Grades by gender

```
ggplot(dat, aes(x = sex, y = G3)) +
  geom_boxplot() +
  labs(
    title = "Final Grade (G3) by Sex",
    x = "Sex",
    y = "Final grade (G3)"
  )
```



Grades by study time:

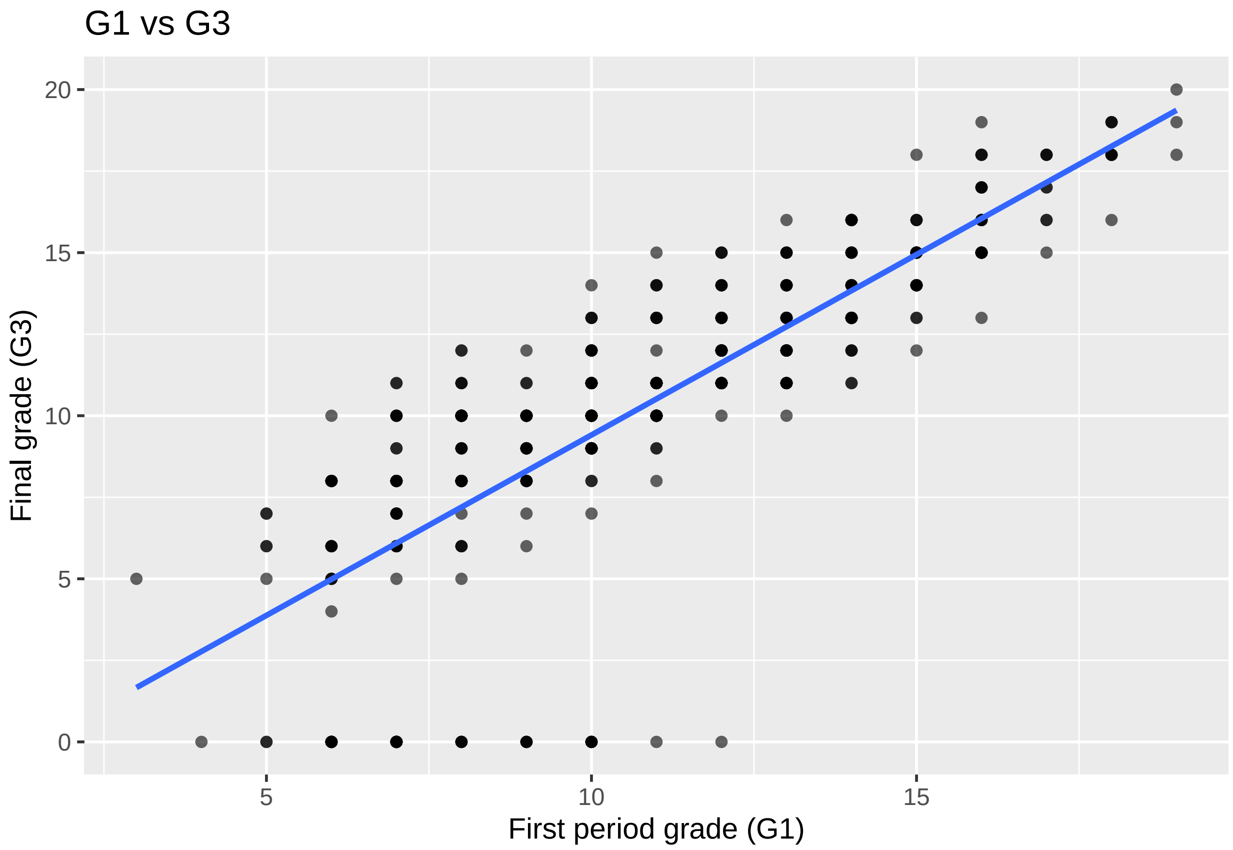
```
ggplot(dat, aes(x = studytime, y = G3)) +
  geom_boxplot() +
  labs(
    title = "Final Grade (G3) by Study Time Category",
    x = "Weekly study time",
    y = "Final grade (G3)"
  )
```



Relationship between grades at different time points

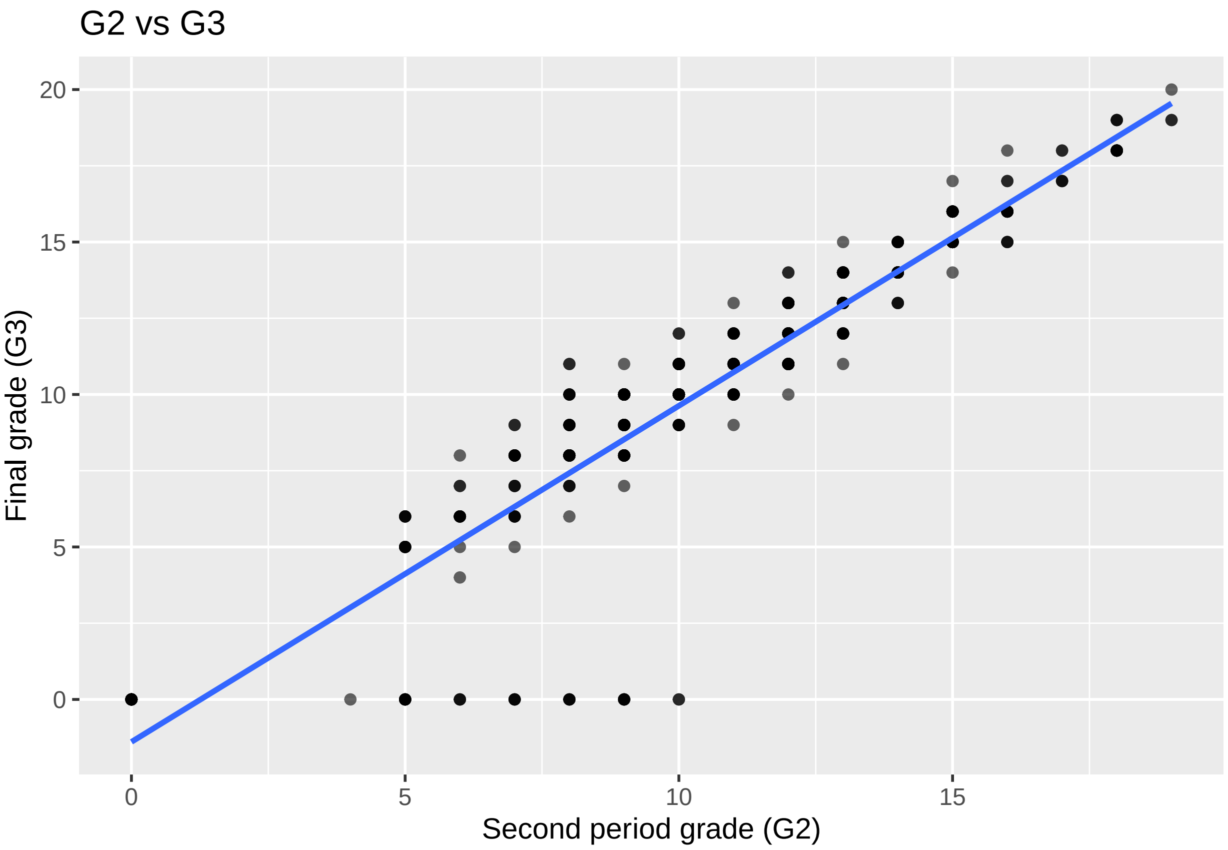
```
#g1
ggplot(dat, aes(x = G1, y = G3)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "G1 vs G3",
    x = "First period grade (G1)",
    y = "Final grade (G3)"
  )
```

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



```
#g2
ggplot(dat, aes(x = G2, y = G3)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "G2 vs G3",
    x = "Second period grade (G2)",
    y = "Final grade (G3)"
  )
```

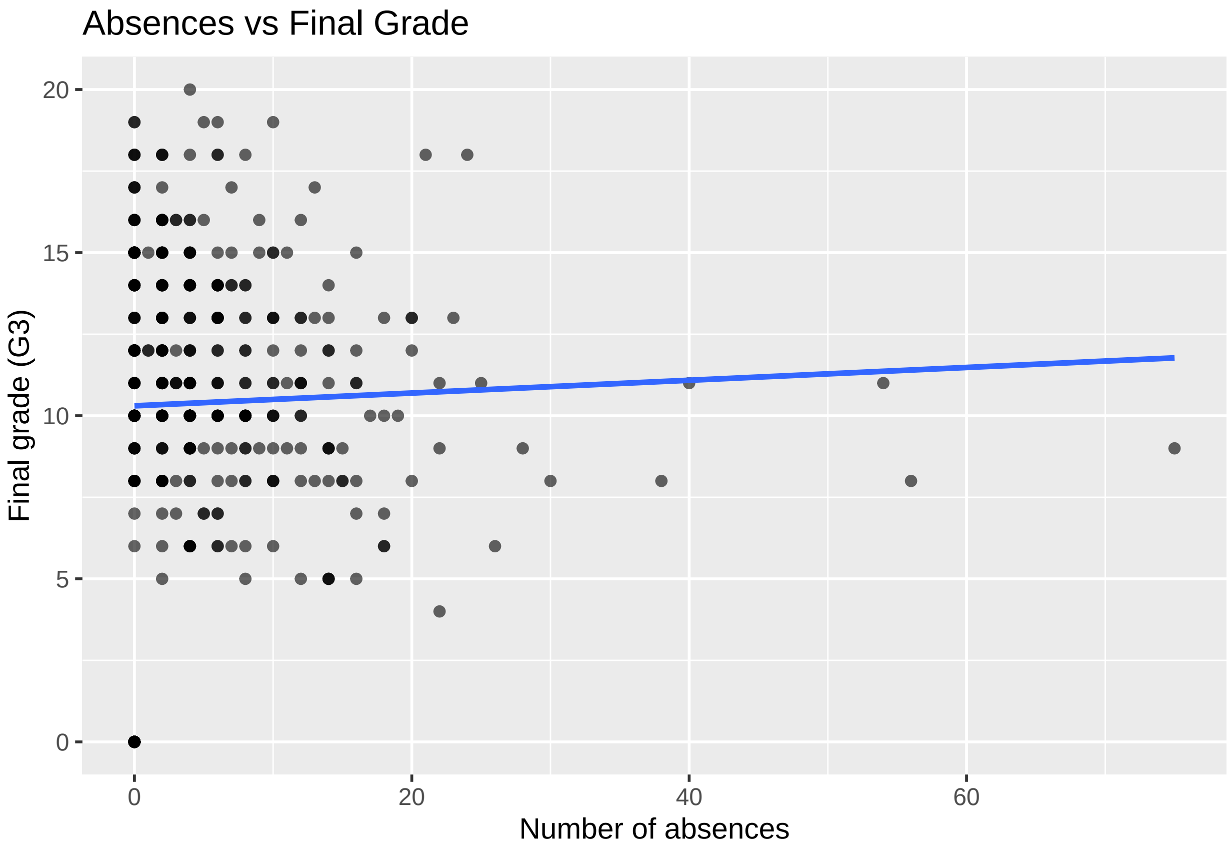
```
## `geom_smooth()` using formula = 'y ~ x'
```



Relationship between absences and final grade:

```r
ggplot(dat, aes(x = absences, y = G3)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Absences vs Final Grade",
    x = "Number of absences",
    y = "Final grade (G3)"
  )
```

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



Some correlations:

```
num_vars <- dat %>%
  select(G1, G2, G3, age, absences)

cor(num_vars)
```

```
##                  G1         G2          G3        age    absences
## G1        1.0000000  0.8521181  0.80146793 -0.0640815 -0.03100290
## G2        0.8521181  1.0000000  0.90486799 -0.1434740 -0.03177670
## G3        0.8014679  0.9048680  1.00000000 -0.1615794  0.03424732
## age      -0.0640815 -0.1434740 -0.16157944  1.0000000  0.17523008
## absences -0.0310029 -0.0317767  0.03424732  0.1752301  1.00000000
```

So correlation is very high between G1/G2/G3, this is expected. Interestingly absences doesn't seem to have much correlation, and neither does age.