modahl_topic_10

Lucas Modahl

2023-04-11

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
d <- read_csv("student-mat.csv") %>% janitor::clean_names()
```

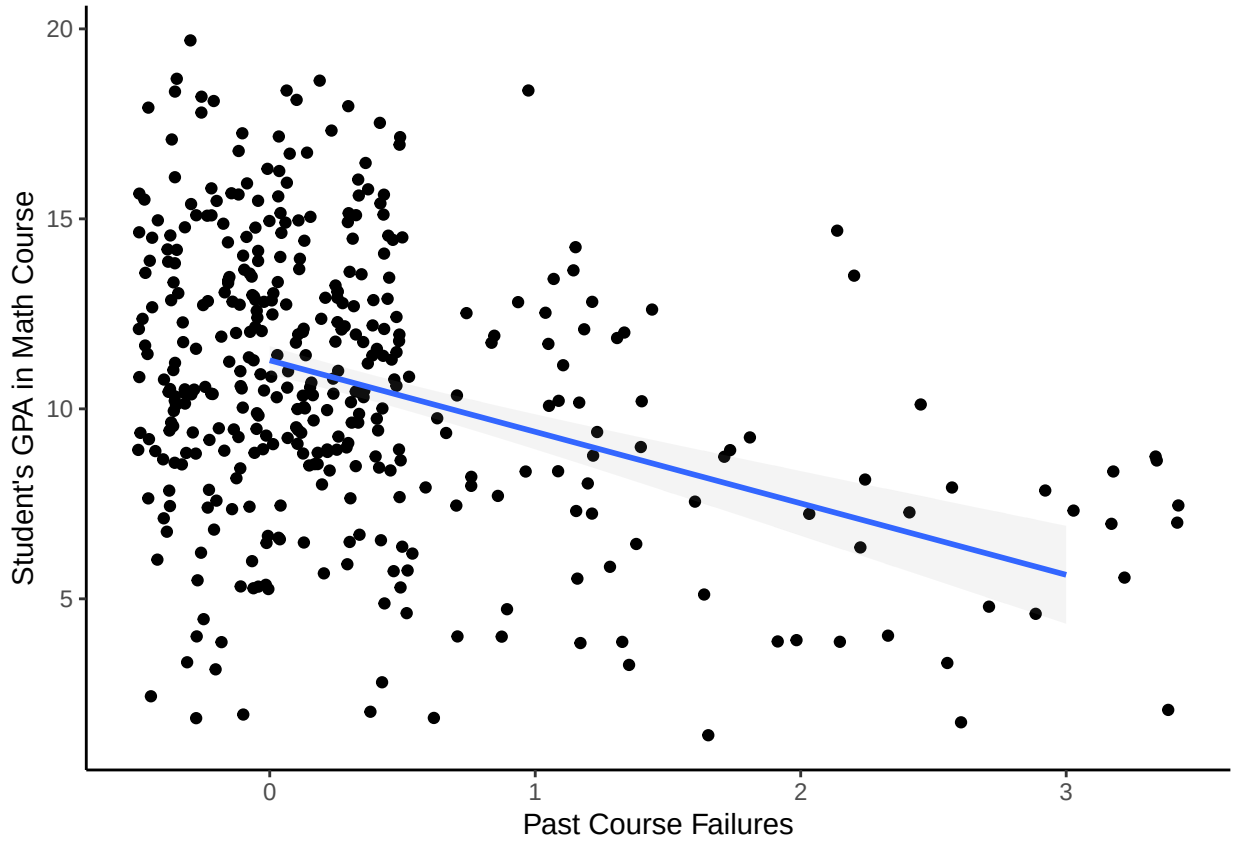
```
## Rows: 395 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...
##
## 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.
```

```
d$gpa <- (d$g1 + d$g2 + d$g3)/3

d$traveltime_str <- dplyr::recode(d$traveltime, "1" = "<15min", "2" = "15-20min", "3" = "30min-1hr", "4

d$studytime_str <- dplyr::recode(d$studytime, "1" = "<2hrs", "2" = "2-5hrs", "3" = "5-10hrs", "4" = ">1

d$goout_str <- dplyr::recode(d$goout, "1" = "very low", "2" = "low", "3" = "moderate", "4" = "high", "5

d <- d[d$medu != 0, ]
d <- d[d$fedu != 0, ]
```

failures

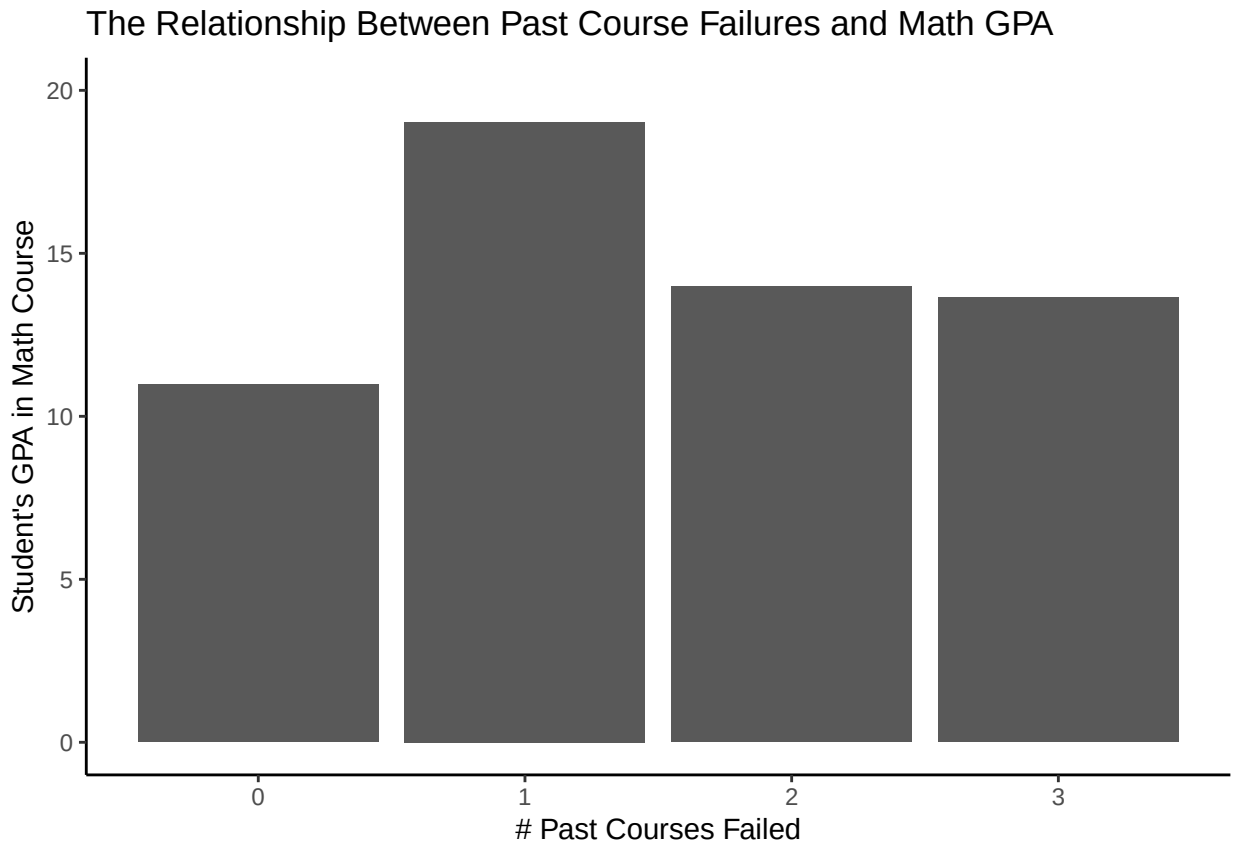
```
d %>%
  ggplot(mapping = aes(x= failures, y = gpa)) +
    geom_jitter(height = .5, width = .5) +
    geom_smooth(method="lm", alpha = .1) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    xlab("Past Course Failures") +
    ylab("Student's GPA in Math Course")
```

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




```
d %>%
  ggplot(mapping = aes(x= failures, y = gpa, group = failures)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
  panel.background = element_blank(), axis.line = element_line(colour = "black")) +
  xlab("# Past Courses Failed") +
  ylab("Student's GPA in Math Course") +
  ggtitle("The Relationship Between Past Course Failures and Math GPA")
```

```
## Warning: Removed 382 rows containing missing values (`geom_bar()`).
```




```
d$absences_c <- d$absences - mean(d$absences)
d$dalc_c <- d$dalc - mean(d$dalc)
d$medu_c <- d$medu - mean(d$medu)
d$fedu_c <- d$fedu - mean(d$fedu)
d$failures_c <- d$failures - mean(d$failures)
d$traveltime_c <- d$traveltime - mean(d$traveltime)
m1 <- lm(gpa ~ failures_c, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = gpa ~ failures_c, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2771 -2.2529  0.0562  2.3895  8.6046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6547     0.1740  61.234  < 2e-16 ***
## failures_c   -1.8817     0.2369  -7.942 2.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.436 on 388 degrees of freedom
```

```
## Multiple R-squared:  0.1398, Adjusted R-squared:  0.1376
## F-statistic: 63.08 on 1 and 388 DF,  p-value: 2.172e-14
```

```
data_summary <- function(data, varname, groupnames){
require(plyr)
summary_func <- function(x, col){
c(mean = mean(x[[col]], na.rm=TRUE),
sd = sd(x[[col]], na.rm=TRUE))
}
data_sum<-ddply(data, groupnames, .fun=summary_func,
varname)
data_sum <- rename(data_sum, c("mean" = varname))
return(data_sum)
}

df2 <- data_summary(d, varname="gpa",
groupnames=c("traveltime_str"))
```

```
## Loading required package: plyr
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
df2$traveltime_str=as.factor(df2$traveltime_str)
head(df2)
```

```
##   traveltime_str       gpa       sd
## 1         <15min 10.977778 3.658178
## 2           >1hr  8.541667 2.701778
## 3       15-20min 10.201923 3.693026
## 4      30min-1hr  9.855072 4.116763
```

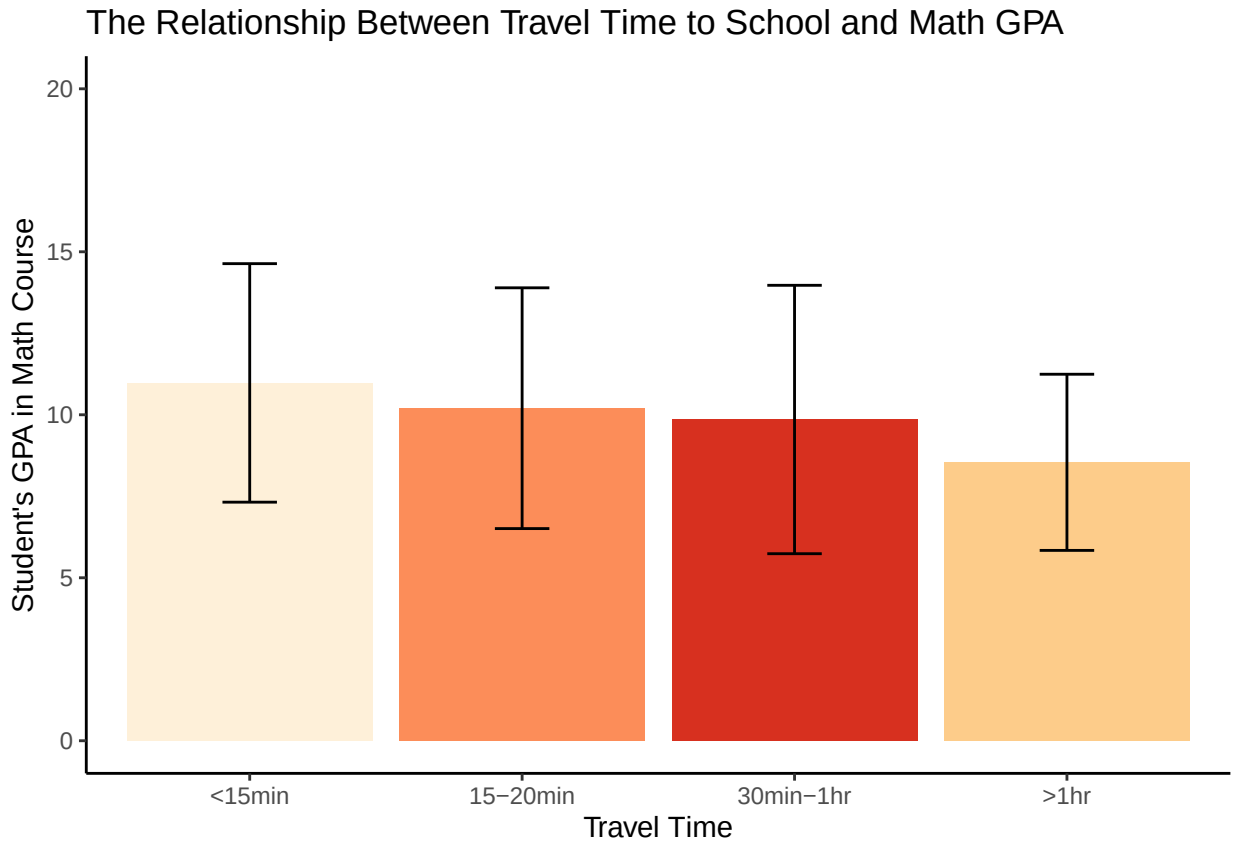
```r
df3 <- data_summary(d, varname="gpa",
groupnames=c("studytime_str"))
df3$studytime_str=as.factor(df3$studytime_str)

df4 <- data_summary(d, varname="gpa",
groupnames=c("schoolsup"))
df4$schoolsup=as.factor(df4$schoolsup)

df5 <- data_summary(d, varname="gpa",
groupnames=c("higher"))
df5$higher=as.factor(df5$higher)

df6 <- data_summary(d, varname="gpa",
groupnames=c("sex"))
df6$sex=as.factor(df6$sex)

df7 <- data_summary(d, varname="gpa",
groupnames=c("address"))
df7$address=as.factor(df7$address)

df8 <- data_summary(d, varname="gpa",
groupnames=c("medu"))
df8$medu=as.factor(df8$medu)

df9 <- data_summary(d, varname="gpa",
groupnames=c("fedu"))
df9$fedu=as.factor(df9$fedu)

df10<- data_summary(d, varname="gpa",
groupnames=c("mjob"))
df10$mjob=as.factor(df10$mjob)

df11<- data_summary(d, varname="gpa",
groupnames=c("internet"))
df11$internet=as.factor(df11$internet)

df12<- data_summary(d, varname="gpa",
groupnames=c("romantic"))
df12$romantic=as.factor(df12$romantic)

df13<- data_summary(d, varname="gpa",
groupnames=c("goout_str"))
df13$goout_str=as.factor(df13$goout_str)
```

```r
head(df4)
```

```
##   schoolsup       gpa       sd
## 1        no 10.849558 3.834167
## 2       yes  9.359477 2.275320
```

travel time

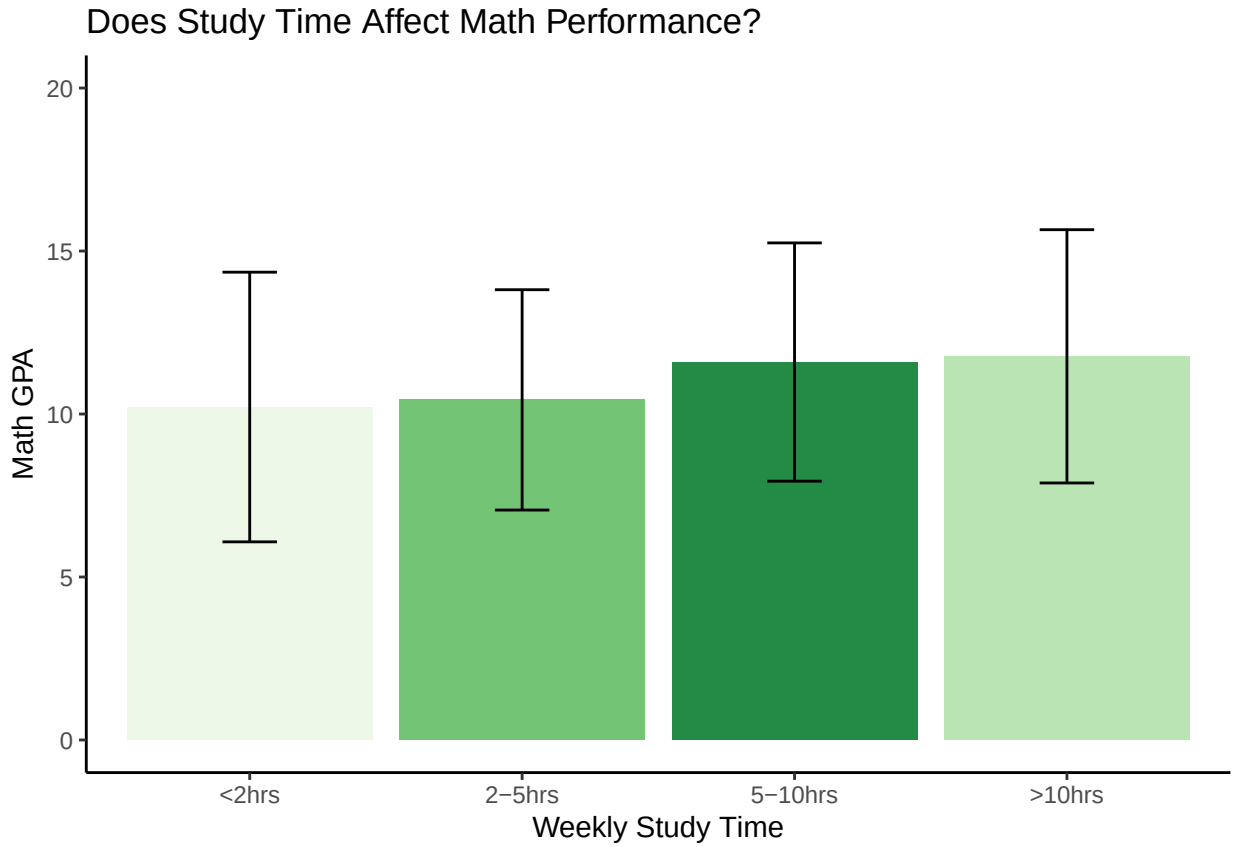
```
df2 %>%
  ggplot(aes(x= reorder(traveltime_str, -gpa), y = gpa, fill = traveltime_str)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Travel Time") +
  ylab("Student's GPA in Math Course") +
  ggtitle("The Relationship Between Travel Time to School and Math GPA") +
  scale_fill_brewer(palette="OrRd") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



study time

```
df3 %>%
  ggplot(aes(x= reorder(studytime_str, gpa), y = gpa, fill = studytime_str)) +
  geom_bar(stat = "identity") +
```

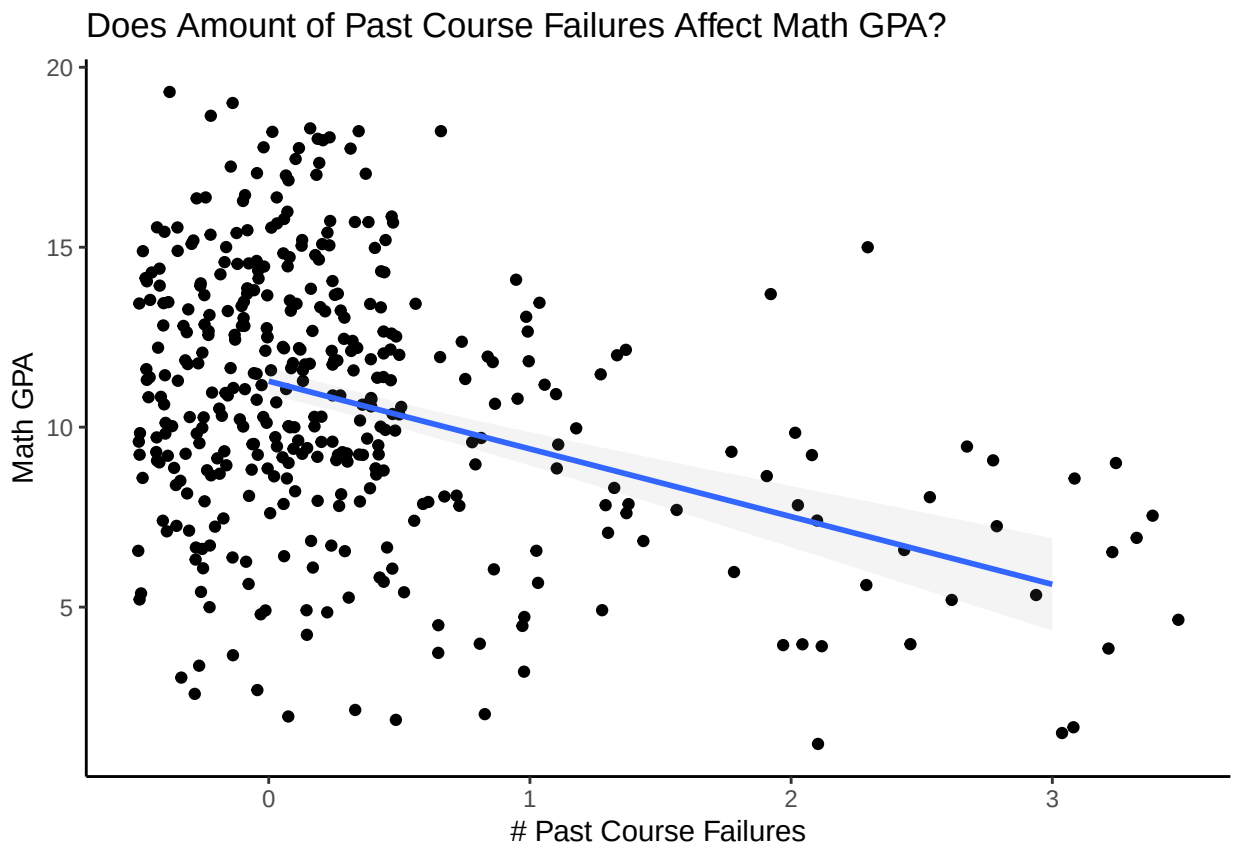
```
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Weekly Study Time") +
  ylab("Math GPA") +
  ggtitle("Does Study Time Affect Math Performance?") +
  scale_fill_brewer(palette="Greens") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



failures

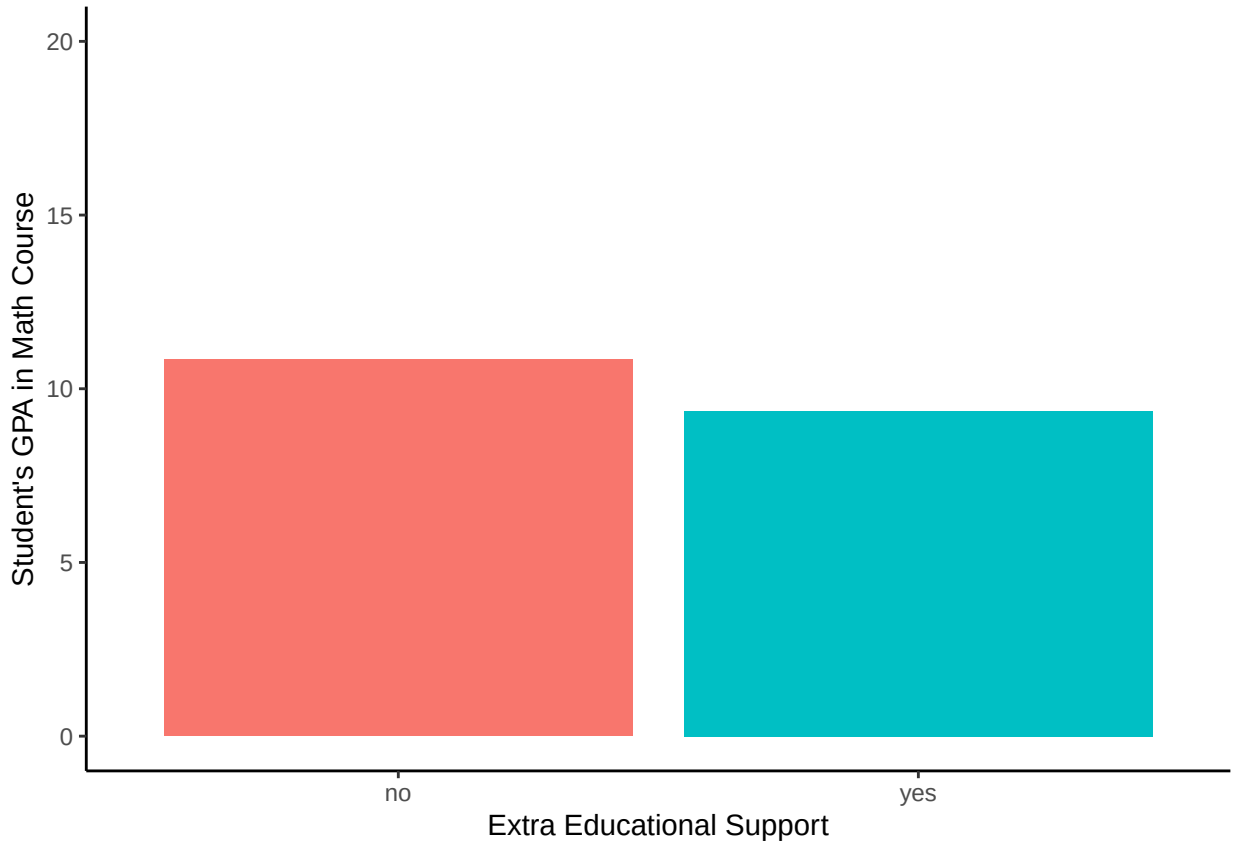
```
d %>%
  ggplot(mapping = aes(x= failures, y = gpa)) +
    geom_jitter(height = .5, width = .5) +
    geom_smooth(method="lm", alpha = .1) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    xlab("# Past Course Failures") +
    ylab("Math GPA") +
  ggtitle("Does Amount of Past Course Failures Affect Math GPA?")
```

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



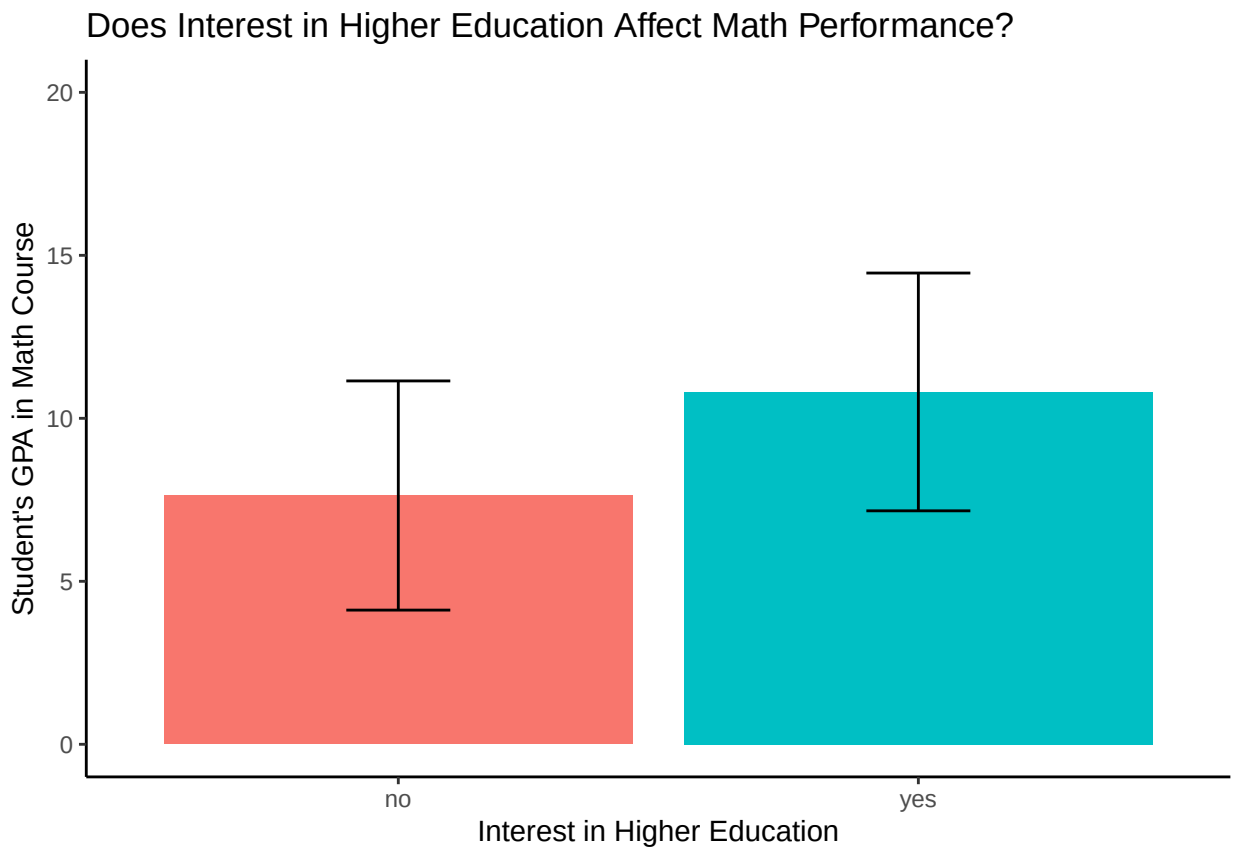
extra edu support

```
df4 %>%
  ggplot(aes(x= schoolsup, y = gpa, fill = schoolsup)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  theme(legend.position="none") +
  xlab("Extra Educational Support") +
  ylab("Student's GPA in Math Course") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



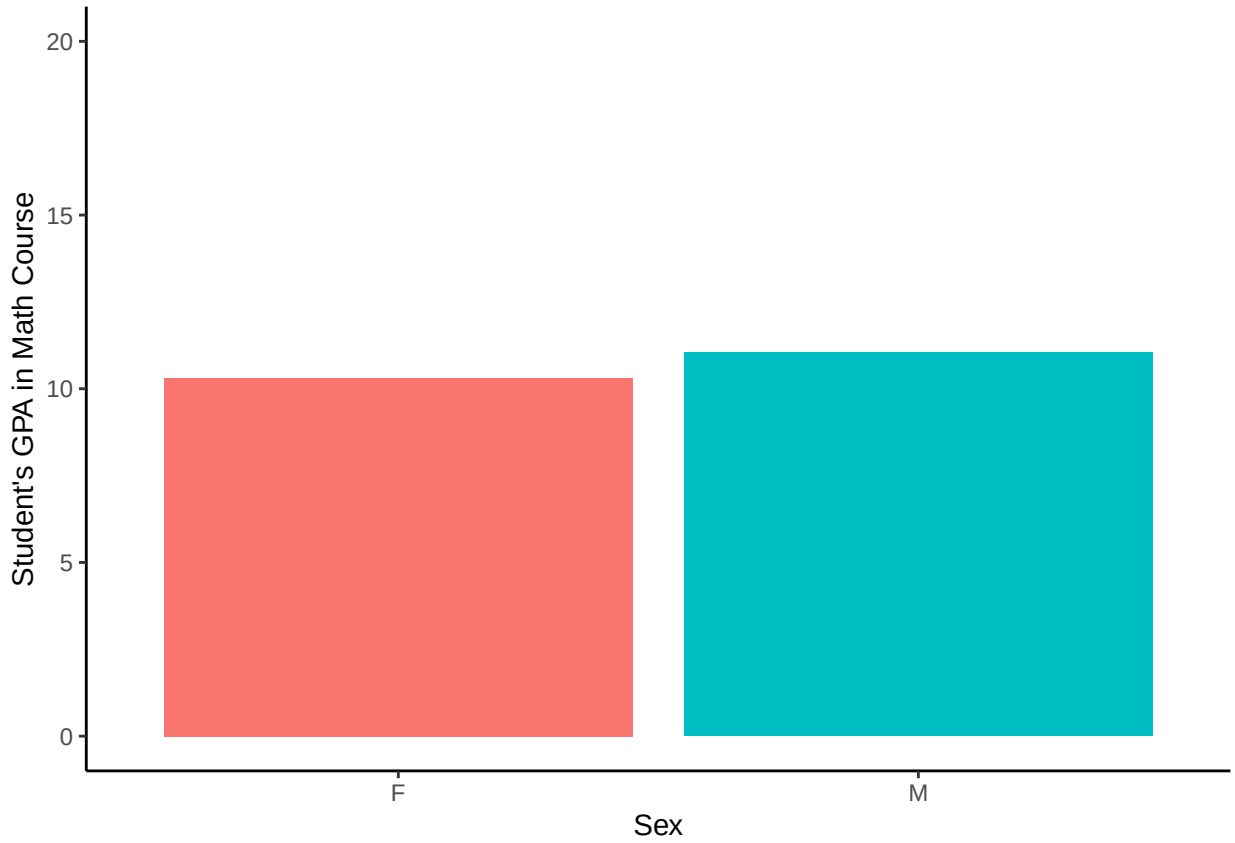
higher

```
df5 %>%
  ggplot(aes(x= higher, y = gpa, fill = higher)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  ylim(0, 20) +
  theme(legend.position="none") +
  xlab("Interest in Higher Education") +
  ylab("Student's GPA in Math Course") +
  ggtitle("Does Interest in Higher Education Affect Math Performance?") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



sex

```
df6 %>%
  ggplot(aes(x= sex, y = gpa, fill = sex)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  theme(legend.position="none") +
  xlab("Sex") +
  ylab("Student's GPA in Math Course") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```

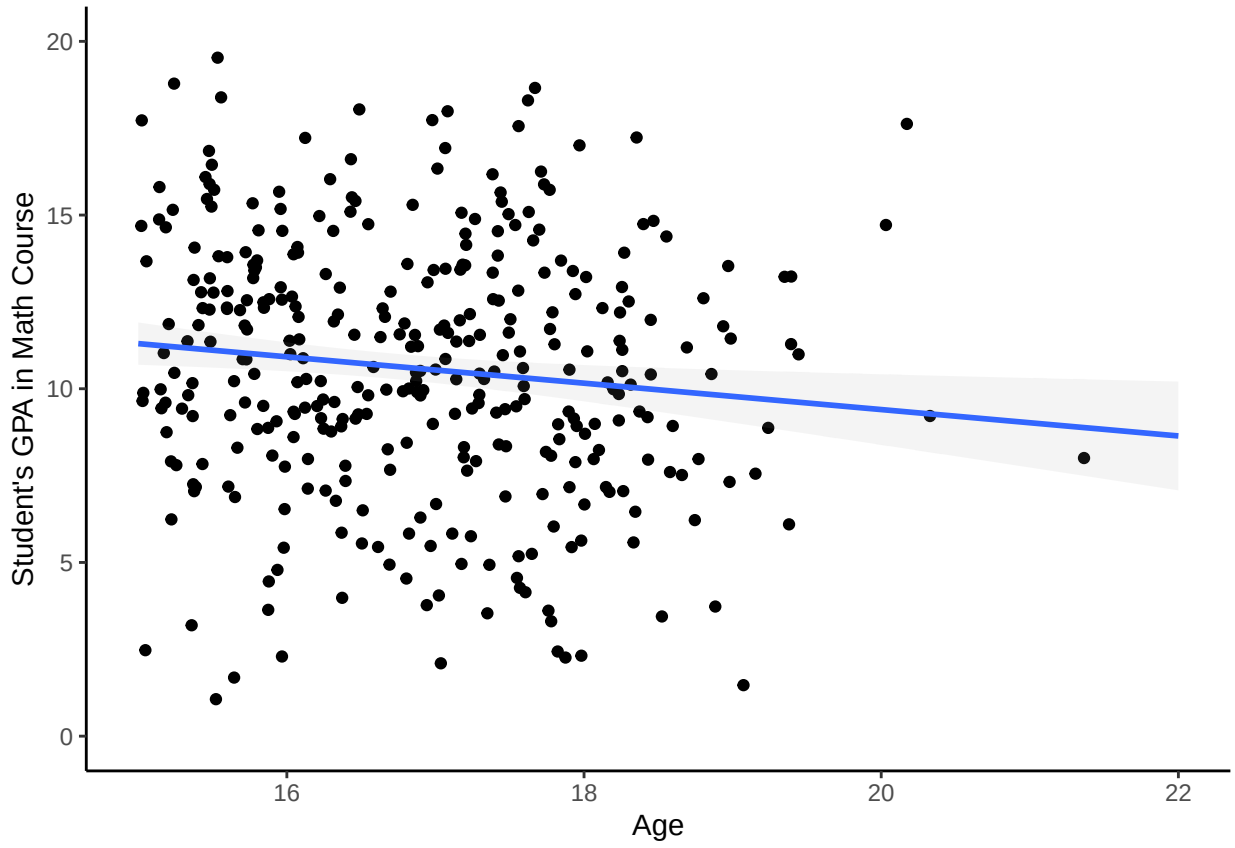


age

```
d %>%
  ggplot(mapping = aes(x= age, y = gpa)) +
    geom_jitter(height = .5, width = .5) +
    geom_smooth(method="lm", alpha = .1) +
    ylim(0, 20) +
    xlim(15, 22) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    xlab("Age") +
    ylab("Student's GPA in Math Course")
```

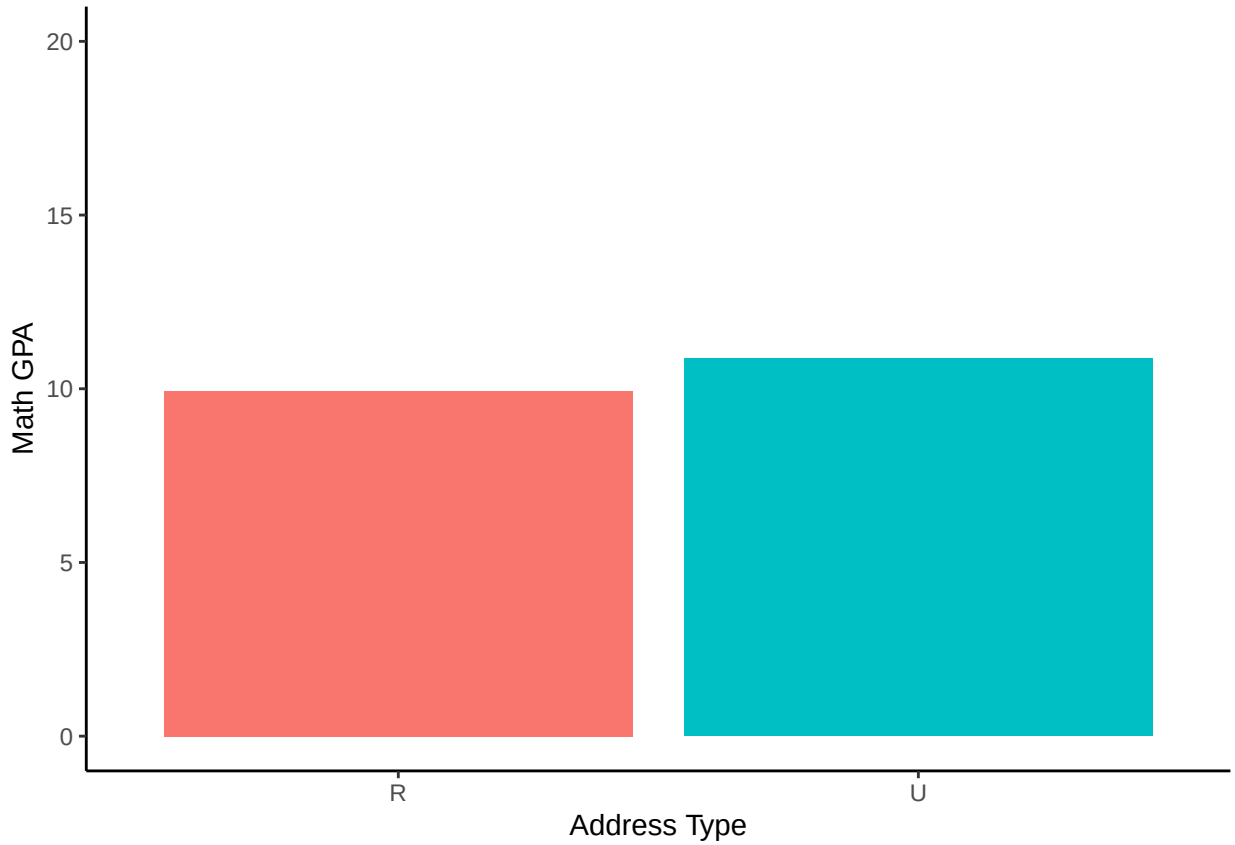
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 37 rows containing missing values (`geom_point()`).
```



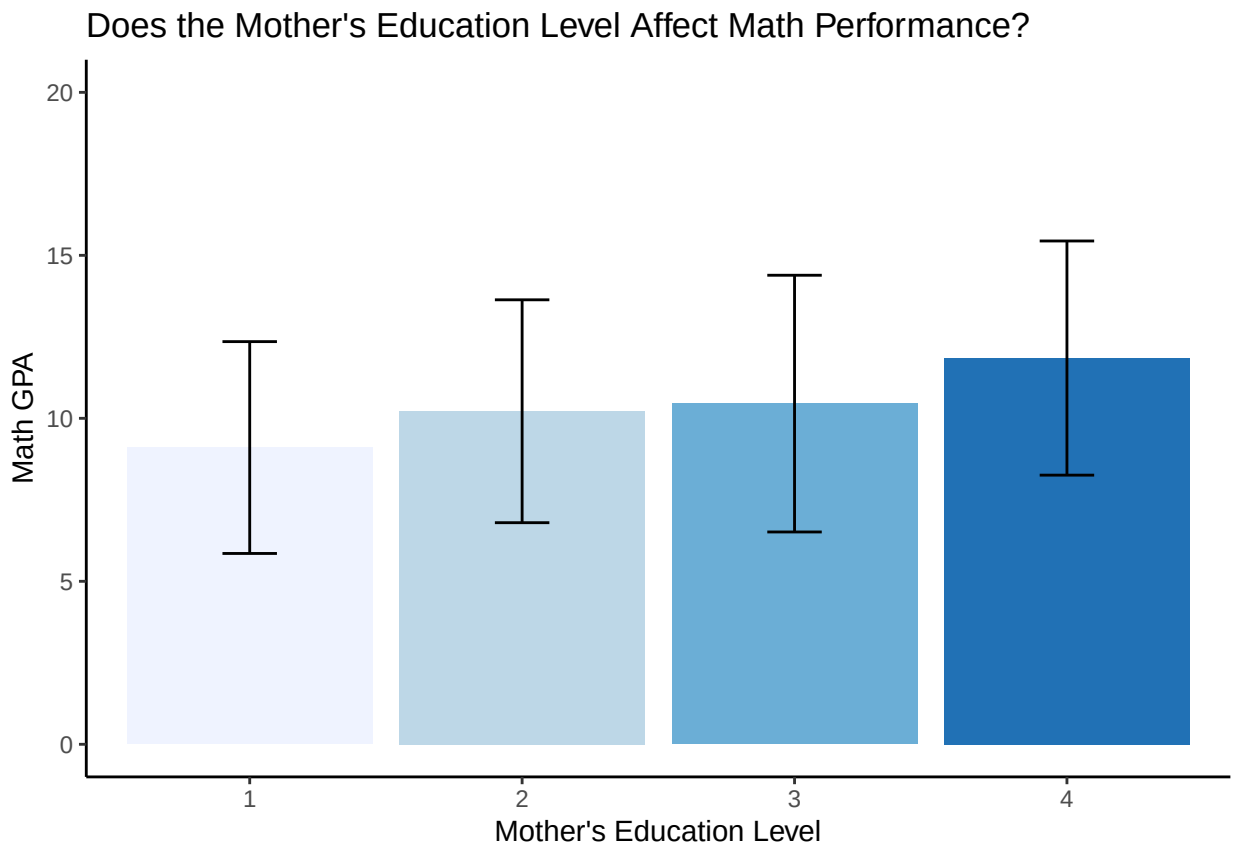
address

```
df7 %>%
  ggplot(aes(x= address, y = gpa, fill = address)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  theme(legend.position="none") +
  xlab("Address Type") +
  ylab("Math GPA") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



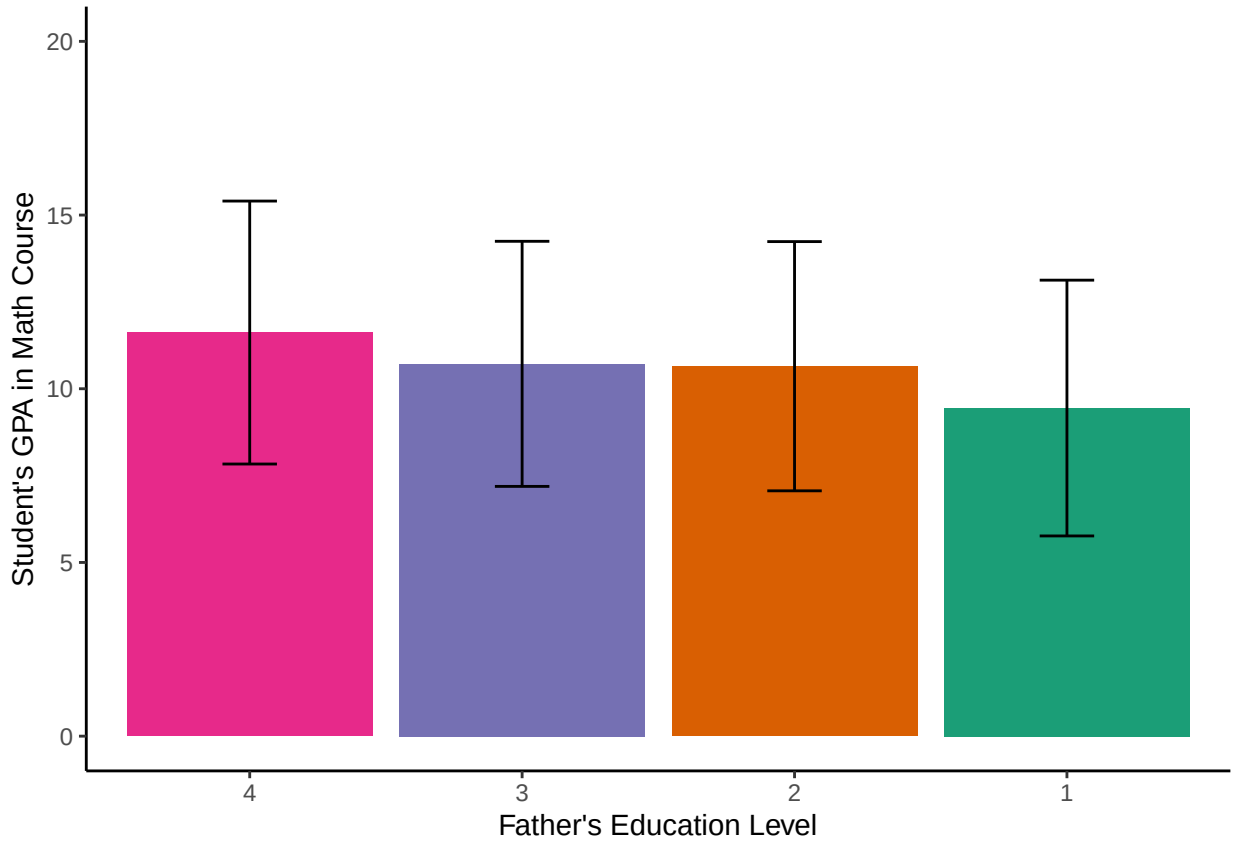
mother's education

```
df8 %>%
  ggplot(aes(x= reorder(medu, gpa), y = gpa, fill = medu)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Mother's Education Level") +
  ylab("Math GPA") +
  scale_fill_brewer(palette="Blues") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black")) +
   ggtitle("Does the Mother's Education Level Affect Math Performance?")
```



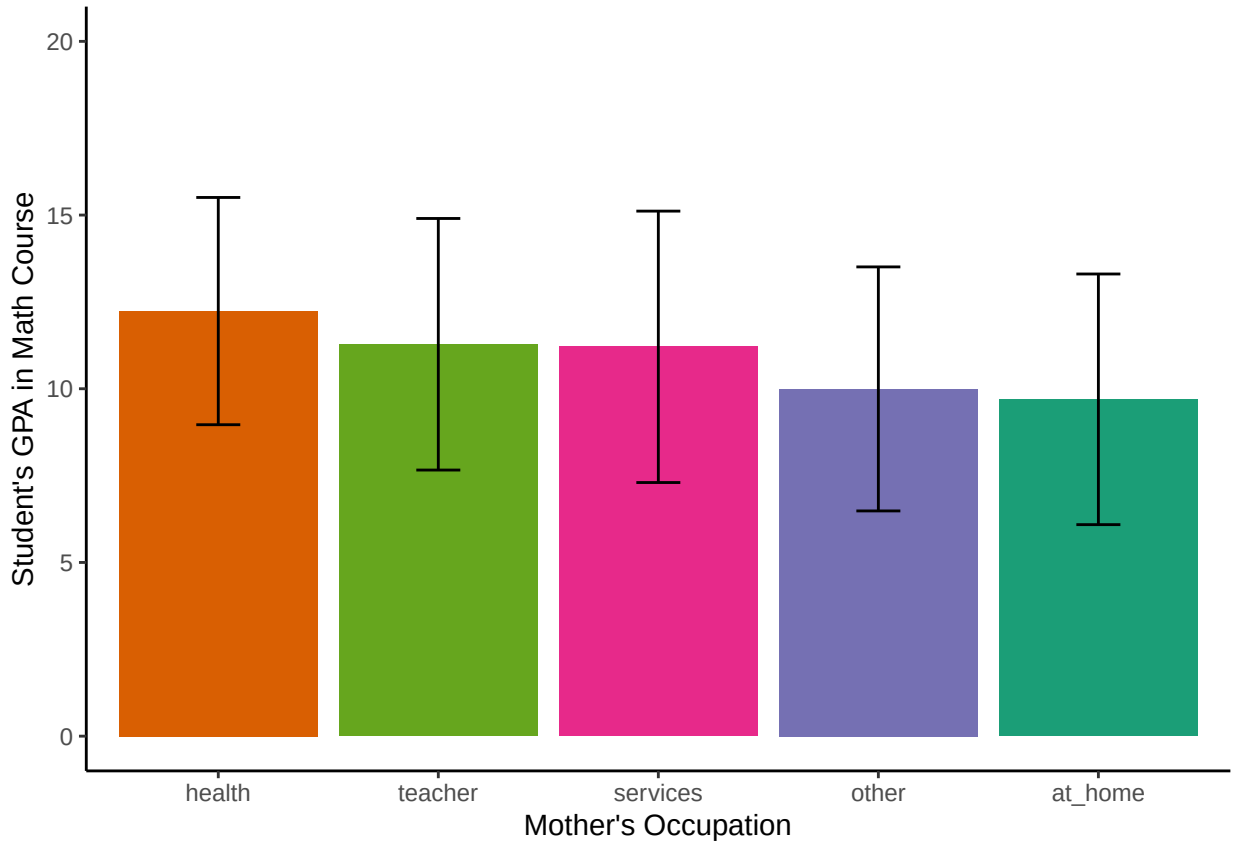
father's education level

```
df9 %>%
  ggplot(aes(x= reorder(fedu, -gpa), y = gpa, fill = fedu)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Father's Education Level") +
  ylab("Student's GPA in Math Course") +
  scale_fill_brewer(palette="Dark2") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



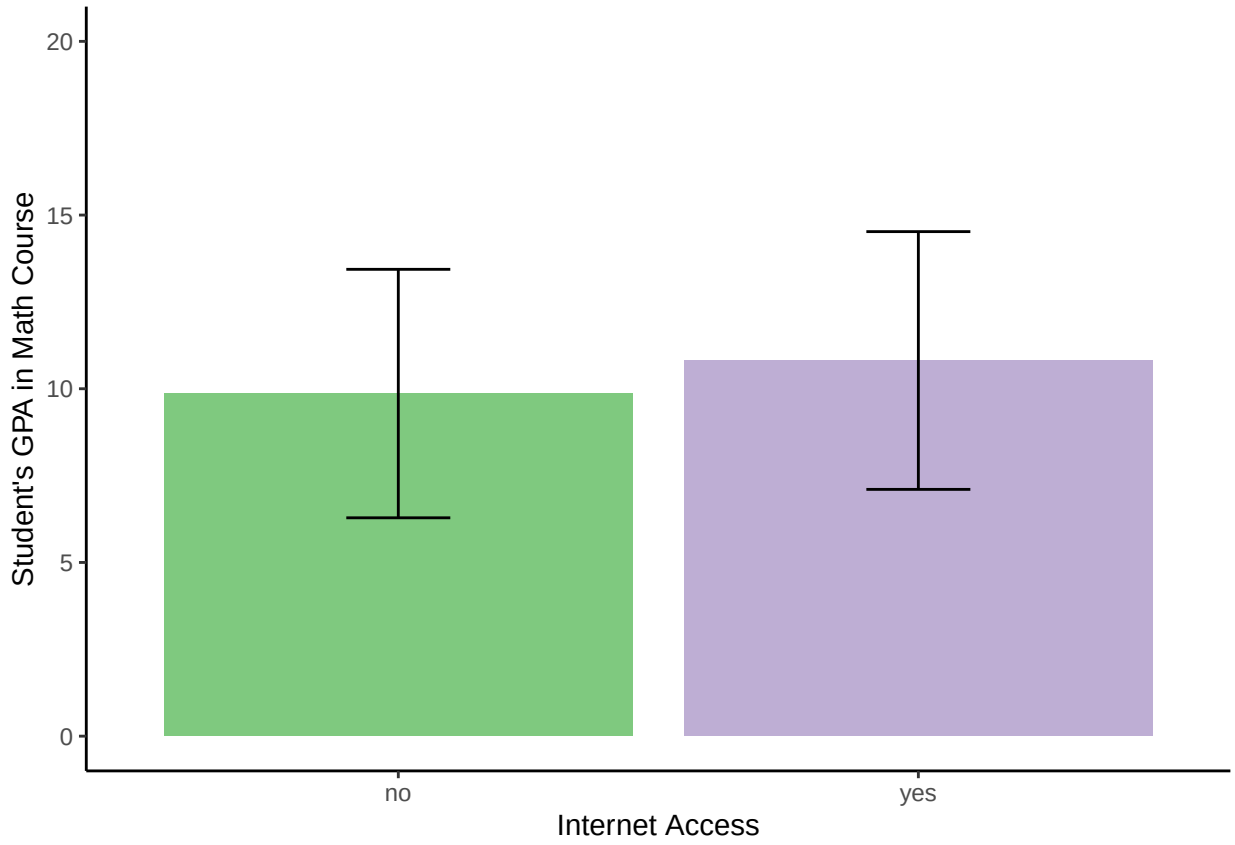
mother's job

```
df10 %>%
  ggplot(aes(x= reorder(mjob, -gpa), y = gpa, fill = mjob)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Mother's Occupation") +
  ylab("Student's GPA in Math Course") +
  scale_fill_brewer(palette="Dark2") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



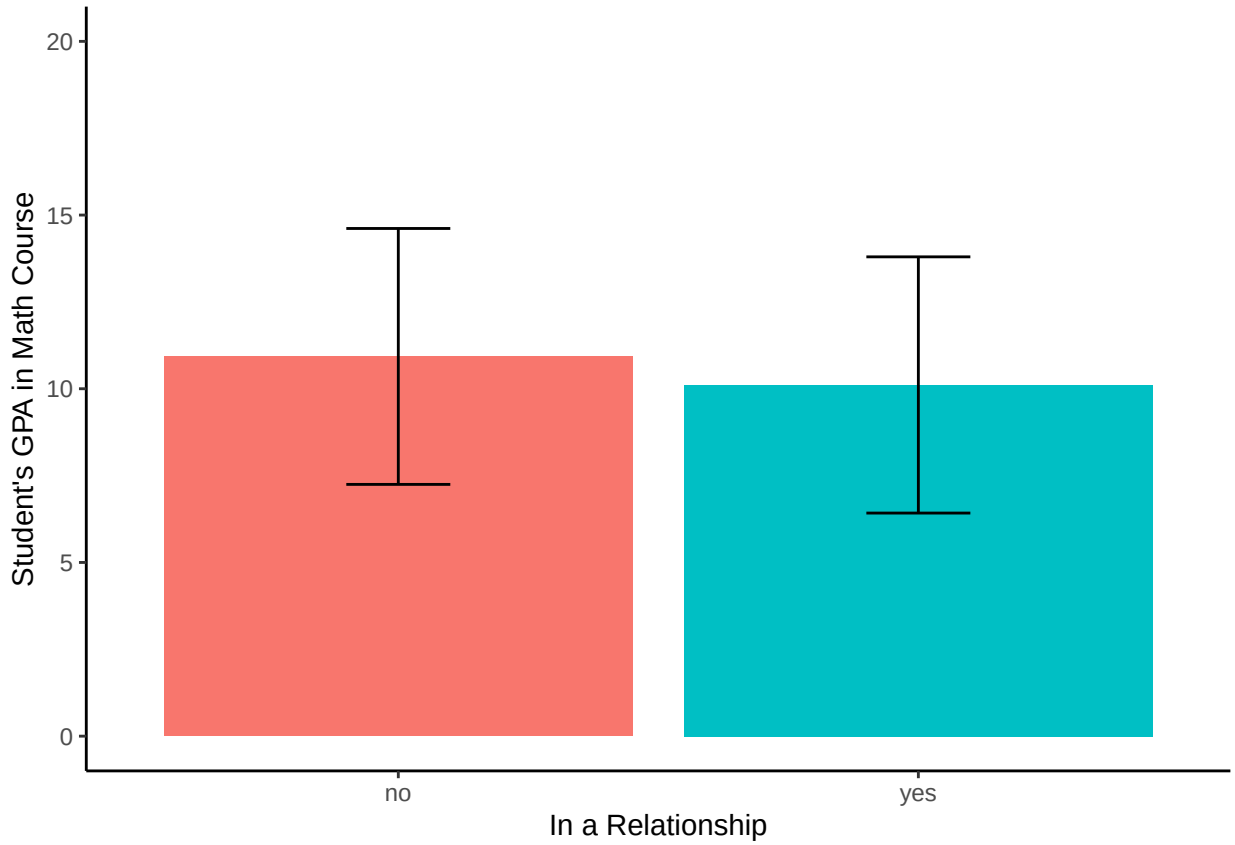
internet access

```
df11 %>%
  ggplot(aes(x= internet, y = gpa, fill = internet)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9))+
  theme(legend.position="none") +
  xlab("Internet Access") +
  ylab("Student's GPA in Math Course") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))+ scale_fill_brewer(
```



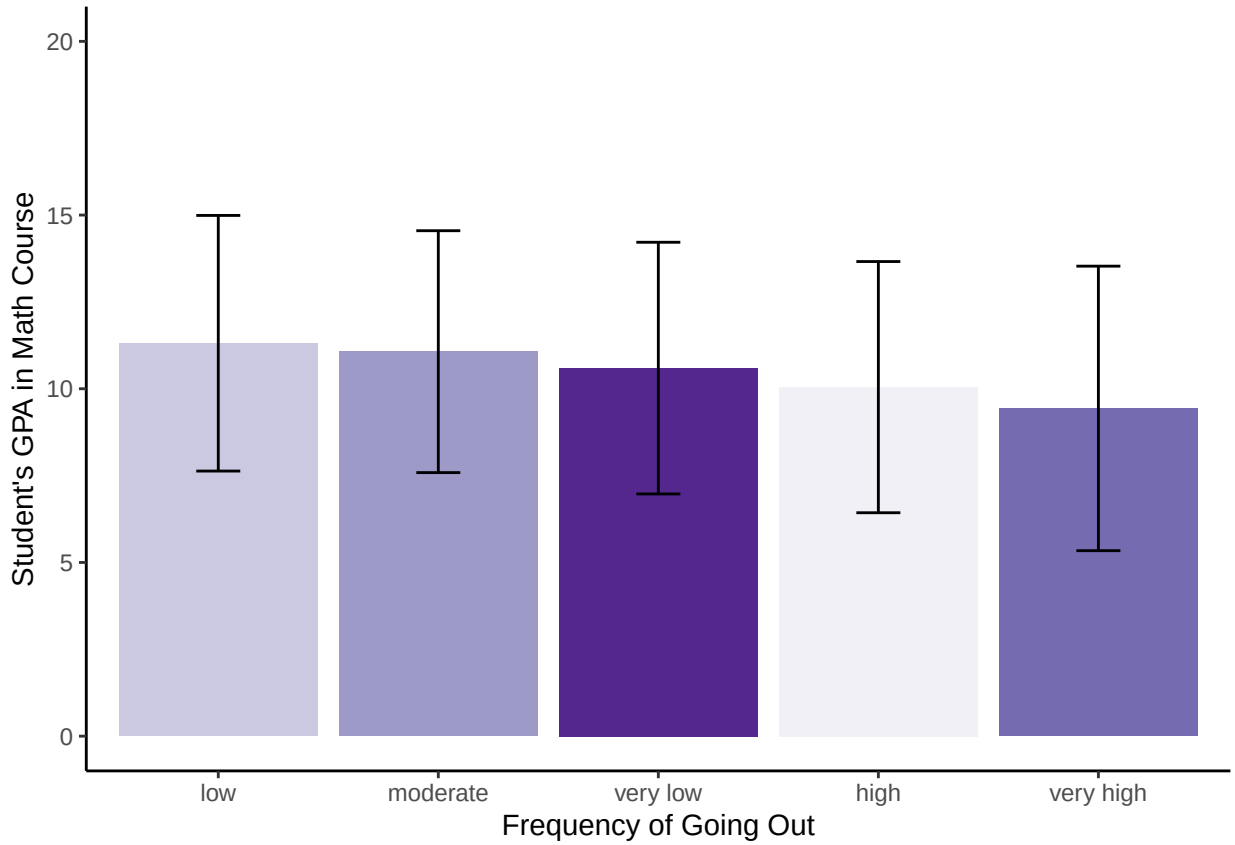
romantic partner

```
df12 %>%
  ggplot(aes(x= romantic, y = gpa, fill = romantic)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9))+
  theme(legend.position="none") +
  xlab("In a Relationship") +
  ylab("Student's GPA in Math Course") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



going out

```
df13 %>%
  ggplot(aes(x= reorder(goout_str, -gpa), y = gpa, fill = goout_str)) +
  geom_bar(stat = "identity") +
  ylim(0, 20) +
  geom_errorbar(aes(ymin=gpa-sd, ymax=gpa+sd), width=.2,
                position=position_dodge(.9)) +
  theme(legend.position="none") +
  xlab("Frequency of Going Out") +
  ylab("Student's GPA in Math Course") +
  scale_fill_brewer(palette="Purples") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    panel.background = element_blank(), axis.line = element_line(colour = "black"))
```




```
head(df8)
```

```
##   medu       gpa       sd
## 1    1  9.103448 3.249356
## 2    2 10.216828 3.418427
## 3    3 10.451178 3.937814
## 4    4 11.848718 3.591576
```

```
d$sex_c <- dplyr::recode(d$sex, "M" = -.5, "F" = .5)
d$address_c <- dplyr::recode(d$address, "R" = -.5, "U" = .5)
d$famsize_c <- dplyr::recode(d$famsize, "LE3" = -.5, "GT3" = .5)
d$pstatus_c <- dplyr::recode(d$pstatus, "A" = -.5, "T" = .5)
d$schoolsup_c <- dplyr::recode(d$schoolsup, "no" = -.5, "yes" = .5)
d$famsup_c <- dplyr::recode(d$famsup, "no" = -.5, "yes" = .5)
d$paid_c <- dplyr::recode(d$paid, "no" = -.5, "yes" = .5)
d$activities_c <- dplyr::recode(d$activities, "no" = -.5, "yes" = .5)
d$nursery_c <- dplyr::recode(d$nursery, "no" = -.5, "yes" = .5)
d$higher_c <- dplyr::recode(d$higher, "no" = -.5, "yes" = .5)
d$internet_c <- dplyr::recode(d$internet, "no" = -.5, "yes" = .5)
d$romantic_c <- dplyr::recode(d$romantic, "no" = -.5, "yes" = .5)
```

```
m1 <- lm(gpa ~ absences, data = d)
summary(m1)
```

```
##
```

```
## Call:
## lm(formula = gpa ~ absences, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3251 -2.3220  0.0099  2.6759  8.6775
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.6583984  0.2307843  46.183   <2e-16 ***
## absences    -0.0006429  0.0233687  -0.028    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.705 on 388 degrees of freedom
## Multiple R-squared:  1.951e-06,  Adjusted R-squared:  -0.002575
## F-statistic: 0.0007569 on 1 and 388 DF,  p-value: 0.9781
```

```
df2 <- data_summary(d, varname="gpa",
groupnames=c("walc"))
df2$walc=as.factor(df2$walc)
head(df2)
```

```
##   walc      gpa       sd
## 1    1 10.98649 4.094477
## 2    2 10.65490 3.822257
## 3    3 10.80169 3.206350
## 4    4  9.72000 3.074554
## 5    5 10.15476 3.337630
```