Statistical Data Analysis of Student Goals

Mateusz Zaremba

November 4, 2019

Contents

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a *document* will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

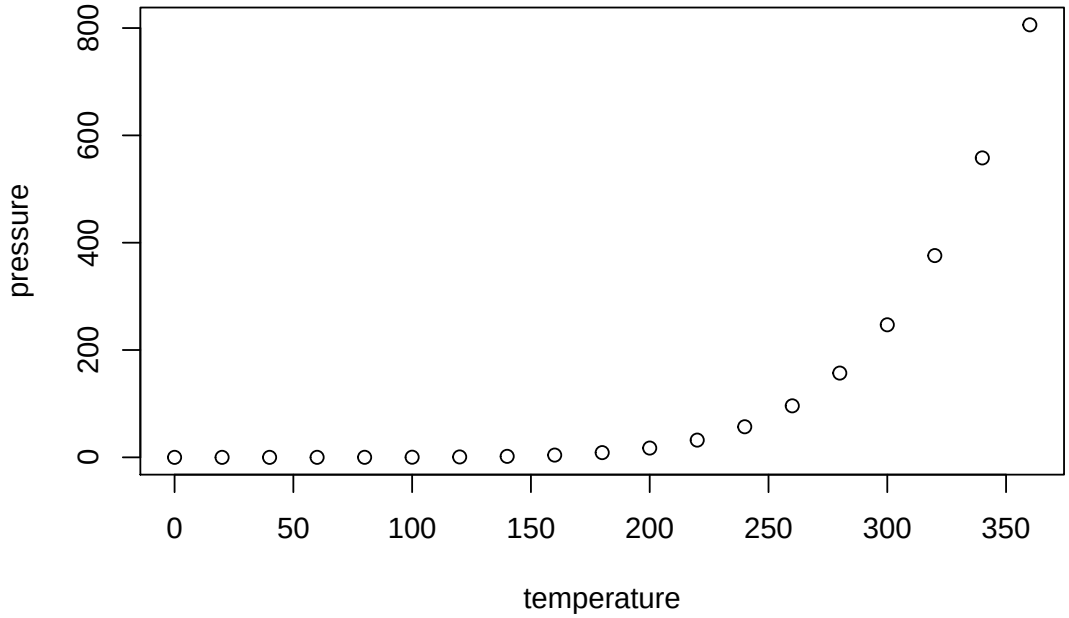
```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:

```
plot(pressure)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
library(rsample)
library(broom)
```

```
##
## Attaching package: 'broom'
```

```
## The following object is masked from 'package:modelr':
##
##     bootstrap
```

```r
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

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

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
# disable warnings for the sake of report writing
options(warn=-1)
```

```r
# set seed for randomization to ensure that results are always reproduced precisely
set.seed(1234)
```

```r
# read csv file (worse variable recognition)
f <- "data/StudentGoalsData.csv"
StudentGoalsData <- read_csv(f)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```r
# drop 'seq' column since it doesn't serve any purpose
StudentGoalsData <- StudentGoalsData  %>%  ungroup  %>%  select(-seq)
```

```r
# count all the students before cleaning and dropping the data
n <- tally(StudentGoalsData)
```

```r
# # clean data - drop results contaiting empty cells
CleanedStudentGoalsData <- drop_na(StudentGoalsData)
```

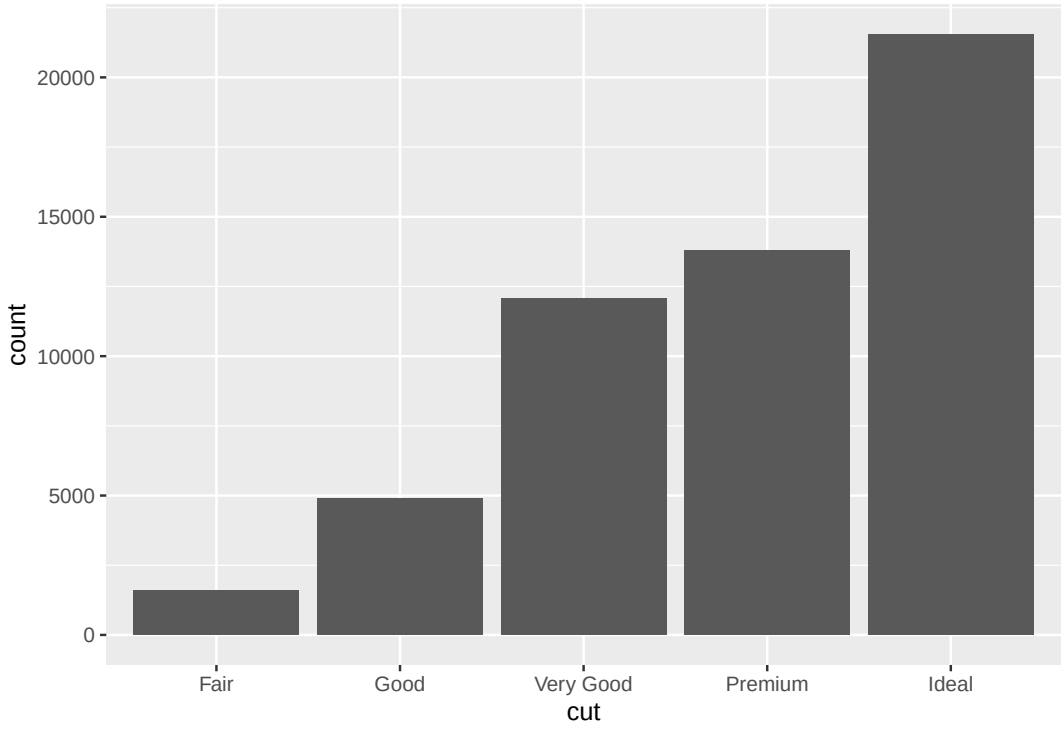
```r
# save CleanedStudentGoalsData table in a simple variable called 'dat'
dat <- CleanedStudentGoalsData
```

```r
# save CleanedStudentGoalsData table as tibble table in a variable called 'dat_tibble'
dat_tibble <- tibble::as_tibble(CleanedStudentGoalsData)
```

```r
# Renaming columns according to random order: 6, 12, 11, 1, 7, 2, 10, 8, 5, 3, 9, 4
renamed_data <- dat_tibble %>%
  rename(
    Q6 = q1,
    Q12 = q2,
    Q11 = q3,
    Q1 = q4,
    Q7 = q5,
    Q2 = q6,
    Q10 = q7,
    Q8 = q8,
    Q5 = q9,
    Q3 = q10,
    Q9 = q11,
    Q4 = q12
  )
# going back to lower case 'q' to keep naming consistent with the original data set
renamed_data2 <- renamed_data %>%
  rename(
    q1 = Q1, q2 = Q2, q3 = Q3, q4 = Q4, q5 = Q5, q6 = Q6,
    q7 = Q7, q8 = Q8, q9 = Q9, q10 = Q10, q11= Q11, q12 = Q12
  )
# save renamed table in 'dat' variable
dat <- renamed_data2
```

```r
# renaming values to their proper labeling from assets/'Student Goals - Coding Information.pdf'
# replace numericals in the 'sex' column with proper sex names
dat$sex[dat$sex==1] <- 'Male'
dat$sex[dat$sex==2] <- 'Female'
# replace numericals in the 'subject' column with proper subject names
dat$subject[dat$subject==1] <- 'Management'
dat$subject[dat$subject==2] <- 'Law'
dat$subject[dat$subject==3] <- 'Tourism'
dat$subject[dat$subject==4] <- 'General Economics'
dat$subject[dat$subject==5] <- 'Accounting'
dat$subject[dat$subject==6] <- 'Statistics'
```

```r
## DATA EXPLORATION #############################################################
# bar chart example
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```



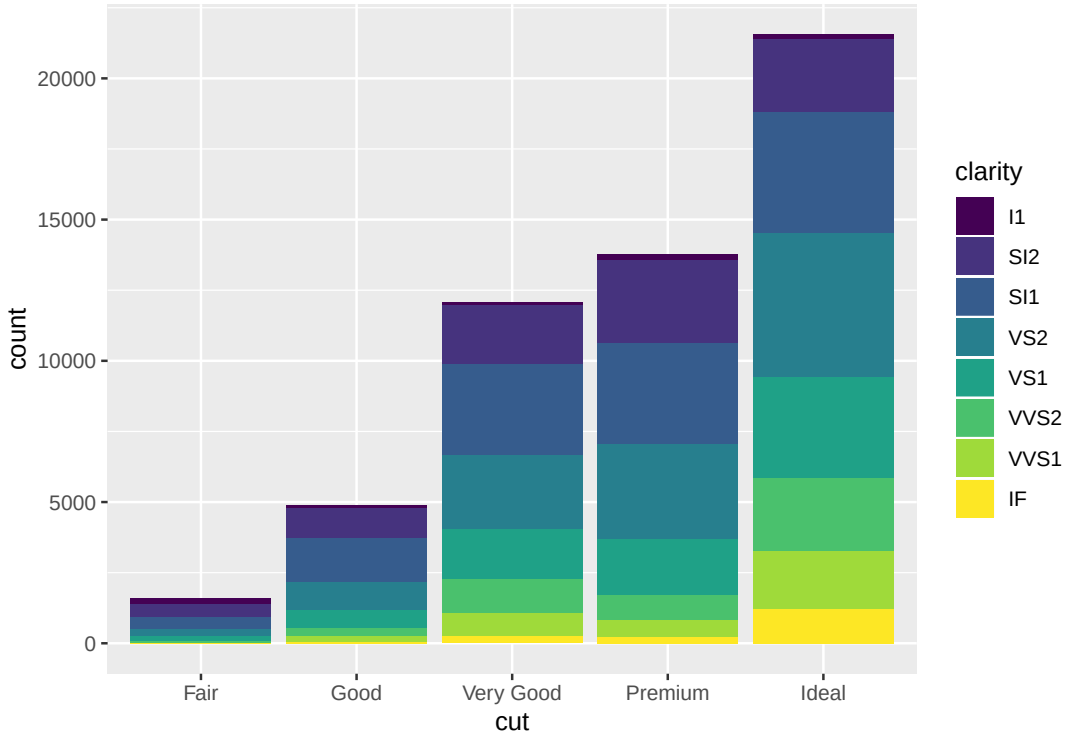

```
head(diamonds)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

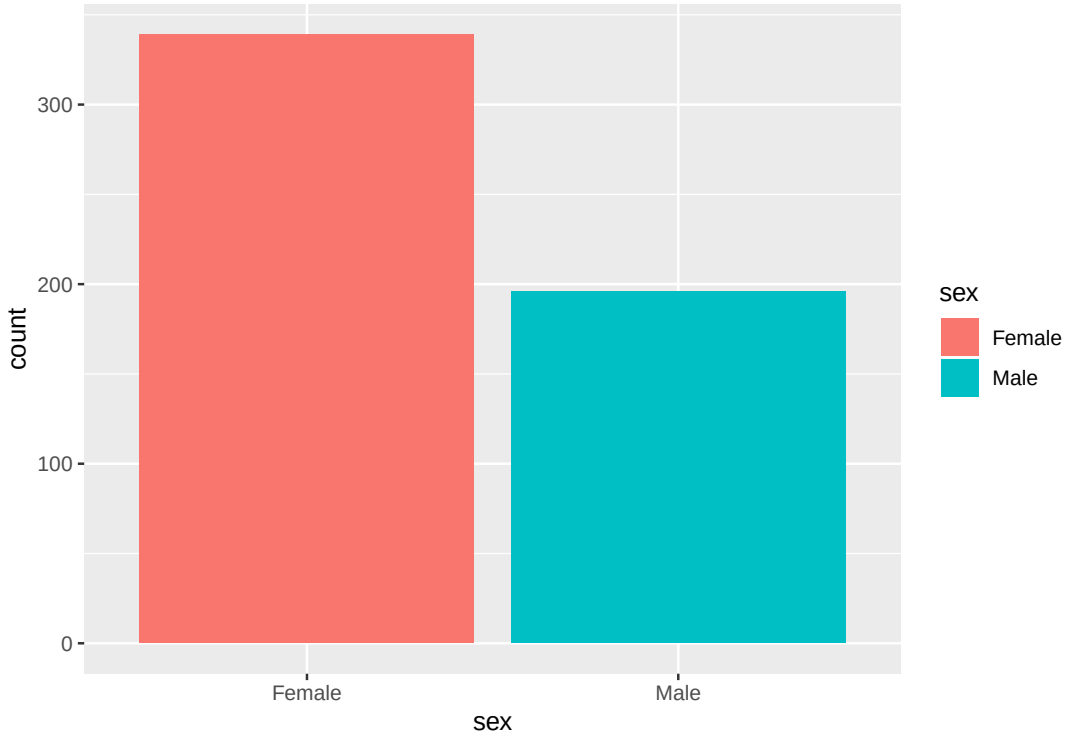
```
head(dat)
```

year	age	sex	subject	q6	q12	q11	q1	q7	q2	q10	q8	q5	q3	q9	q4	interest	enj
4	20	Male	Management	6	2	4	6	7	5	4	5	1	3	6	6	7	
4	20	Male	Management	3	1	1	3	5	3	4	5	3	1	5	1	7	
3	19	Female	Management	2	2	1	6	5	4	1	7	1	1	5	1	7	
3	19	Male	Management	4	5	5	3	6	4	6	6	4	4	6	3	7	
3	18	Female	Management	6	4	2	4	7	4	6	7	2	2	7	2	7	
3	19	Male	Management	7	2	2	6	7	6	7	7	5	5	7	5	7	

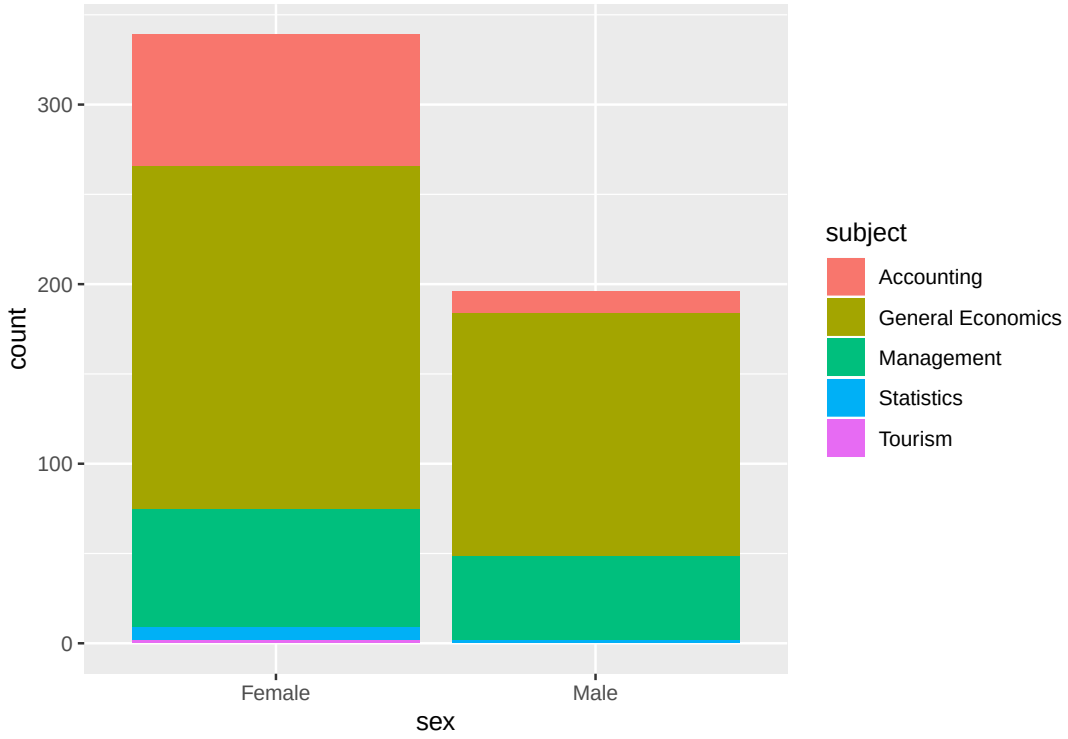
```r
#bar chart example 2
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity))
```



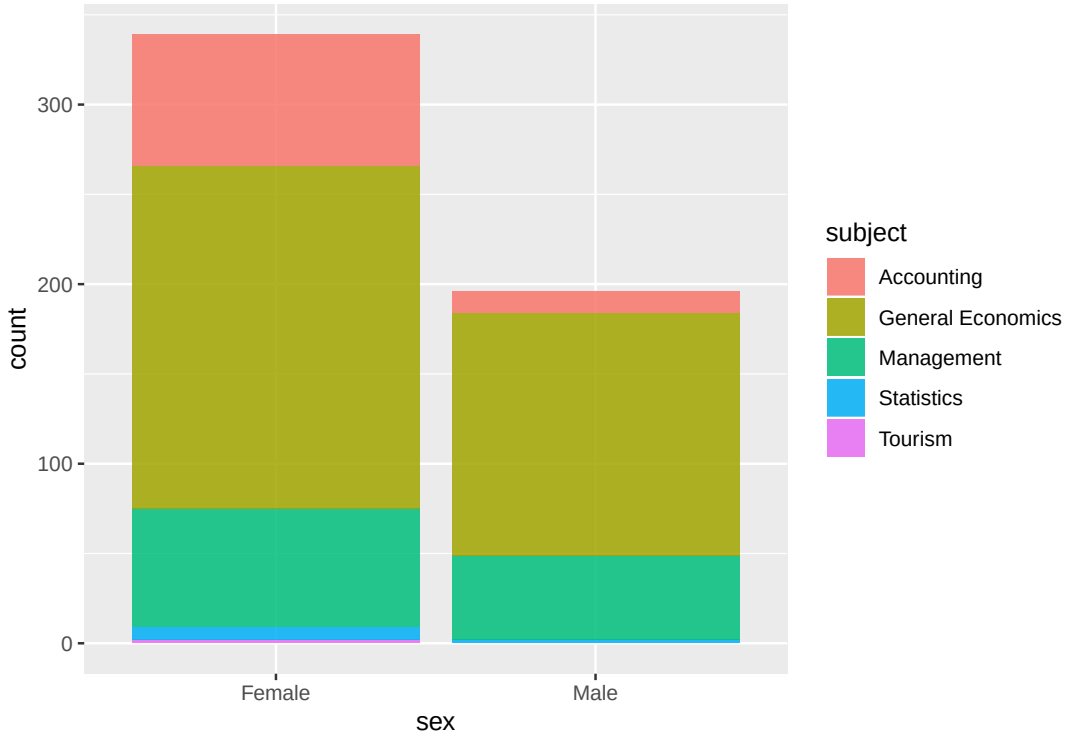

```r
# gender in numbers
ggplot(data = dat) +
  geom_bar(mapping = aes(x = sex, fill = sex))
```



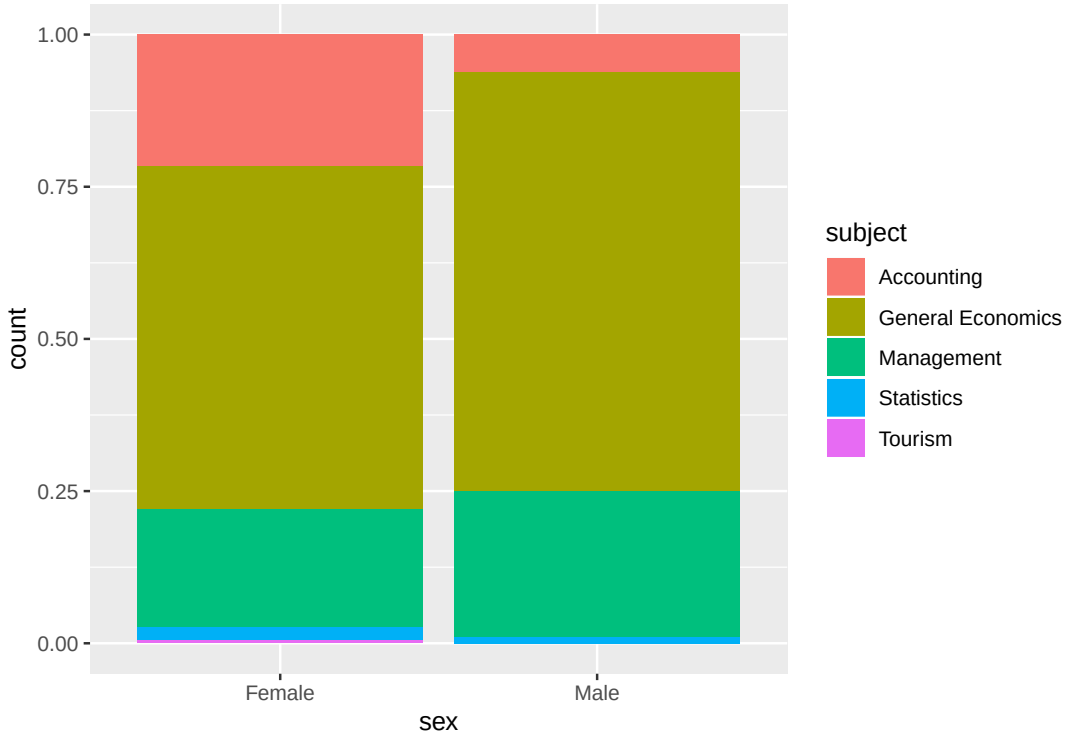

```
# subject by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = sex, fill = subject))
```



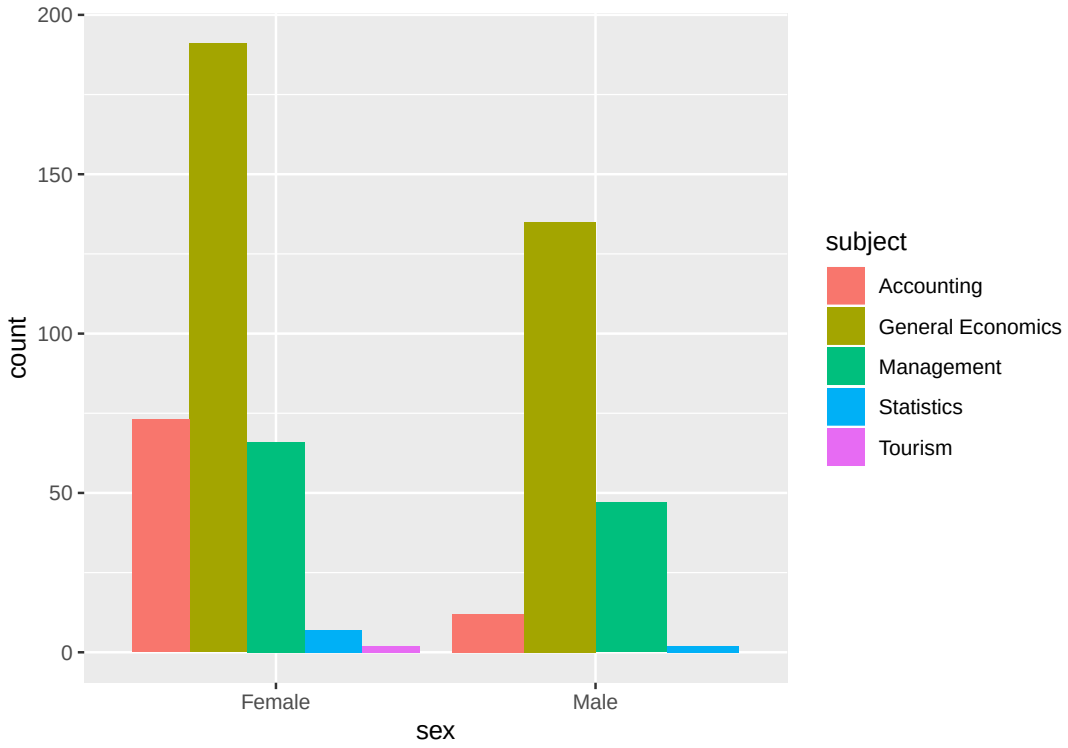

```
#subject by gender with alpha blending
ggplot(data = dat) +
  geom_bar(alpha = 0.85, mapping = aes(x = sex, fill = subject))
```



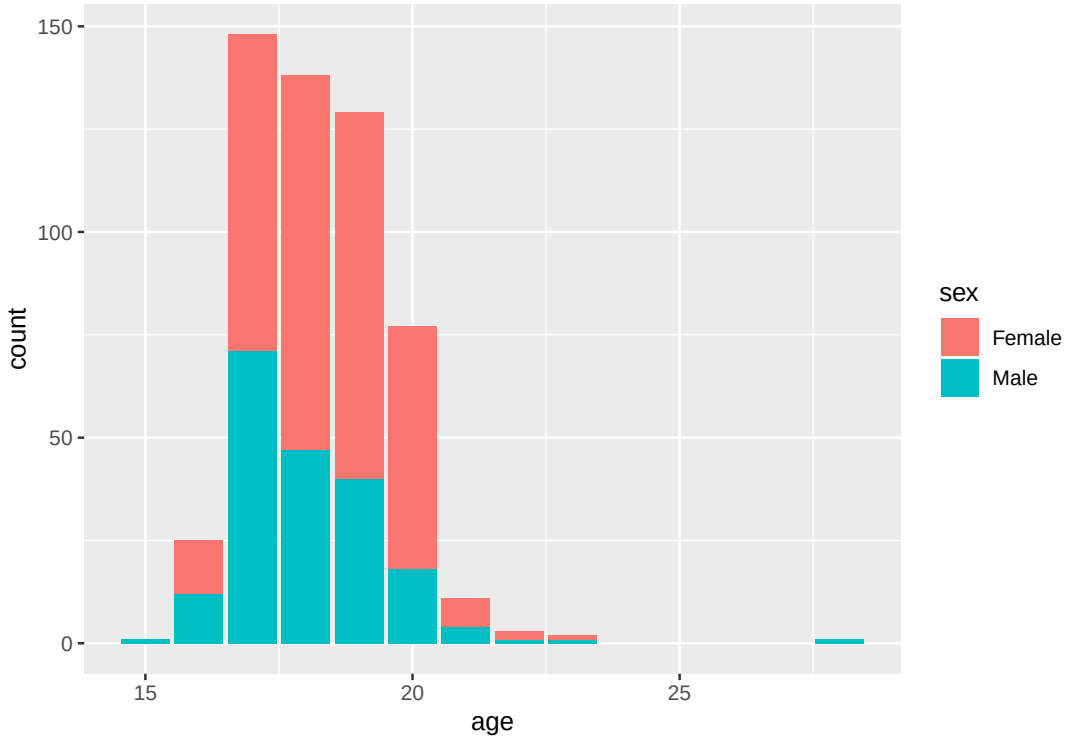

```
#subject by gender and normalizing using position = "fill"
ggplot(data = dat) +
  geom_bar(mapping = aes(x = sex, fill = subject), position = "fill")
```



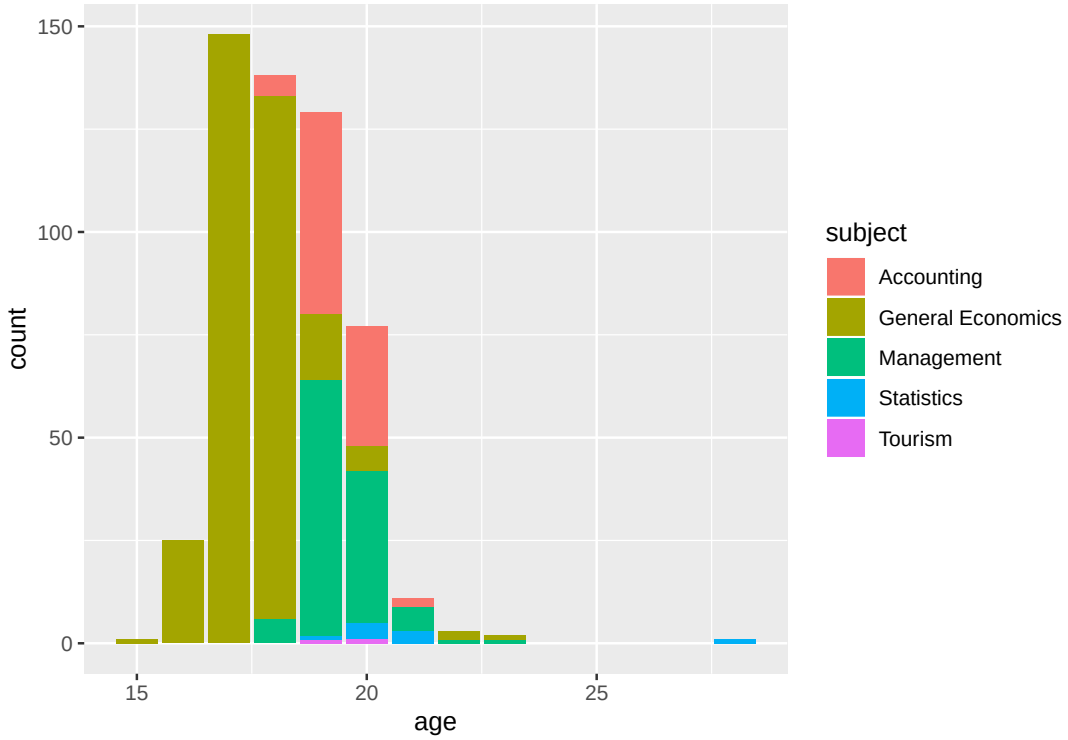

```
#subject by gender and normalizing using position = "dodge" to place overlapping objects directly
ggplot(data = dat) +
  geom_bar(mapping = aes(x = sex, fill = subject), position = "dodge")
```



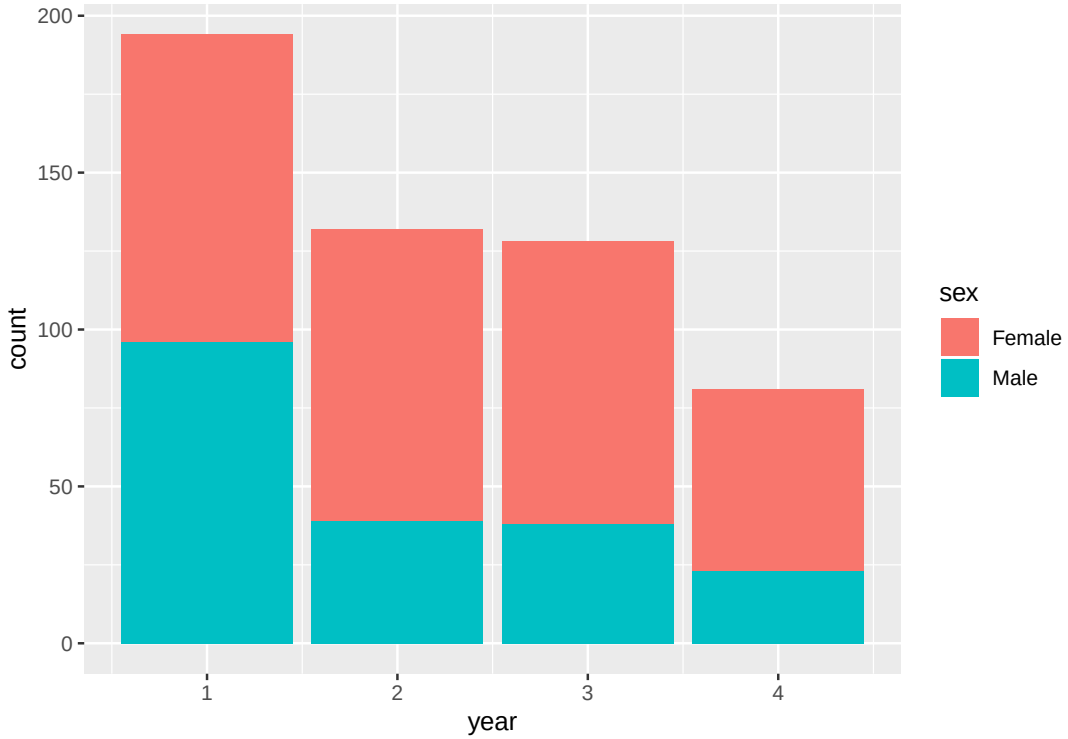

```
# age by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = age, fill = sex))
```



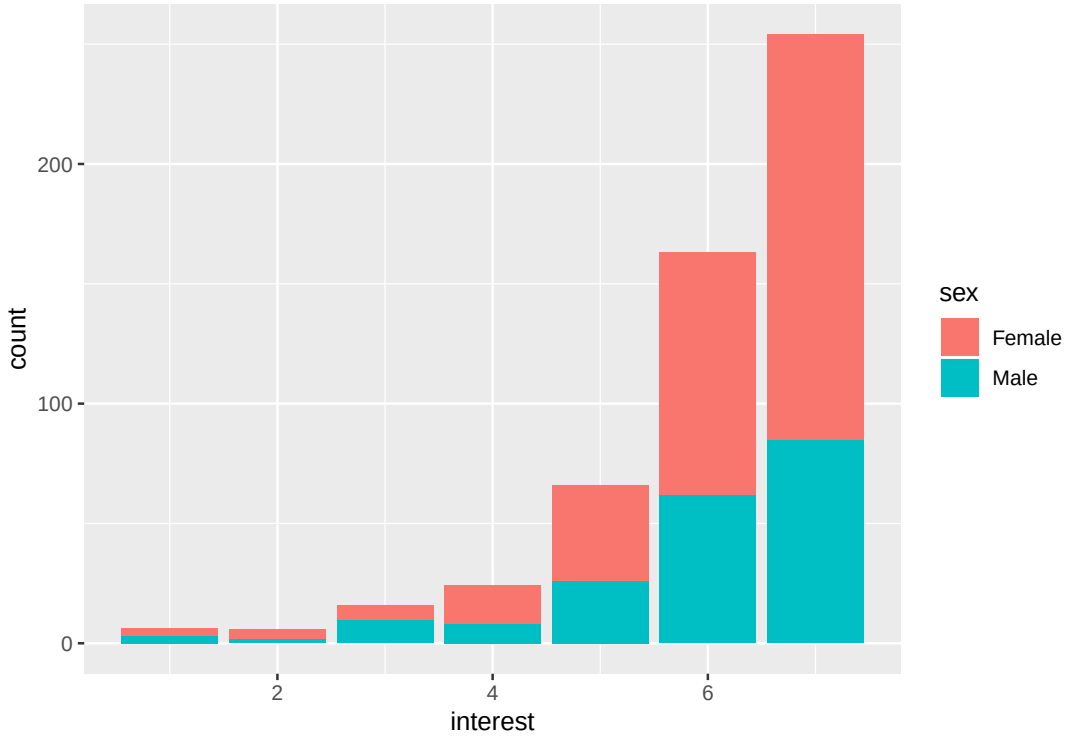

```
# age by subject
ggplot(data = dat) +
  geom_bar(mapping = aes(x = age, fill = subject))
```



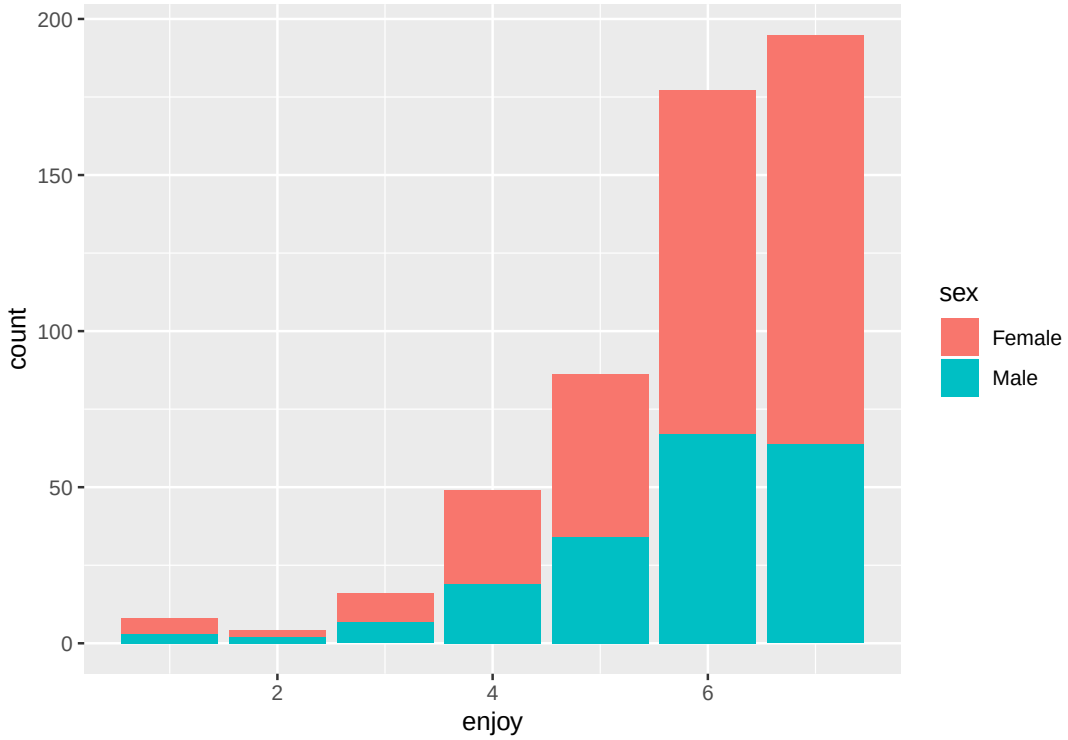

```
# course year by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = year, fill = sex))
```



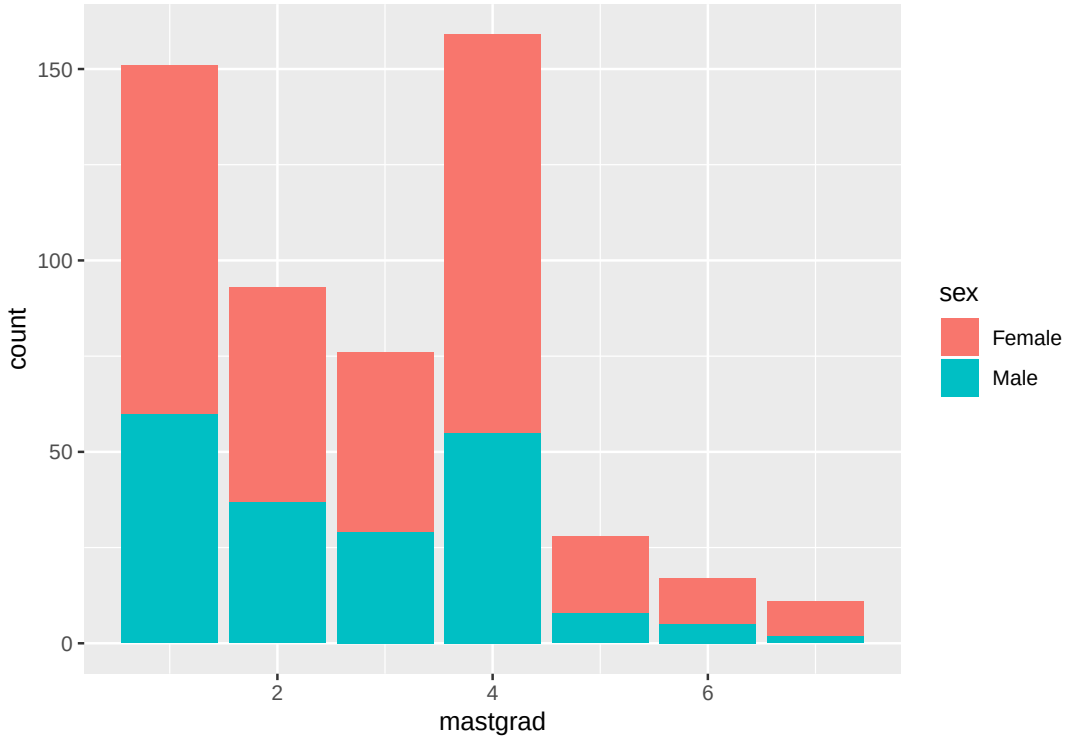

```
# "expect my courses this semester to be very interesting" by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = interest, fill = sex))
```



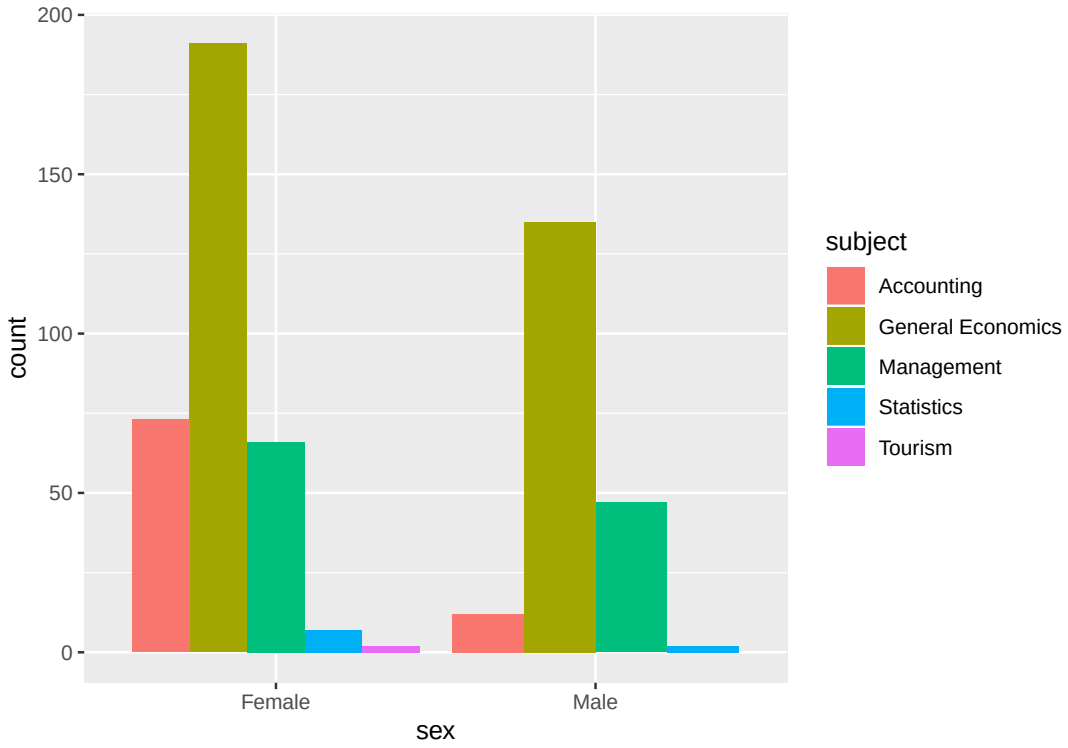

```
# "expect my courses this semester to be very enjoyable" by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = enjoy, fill = sex))
```



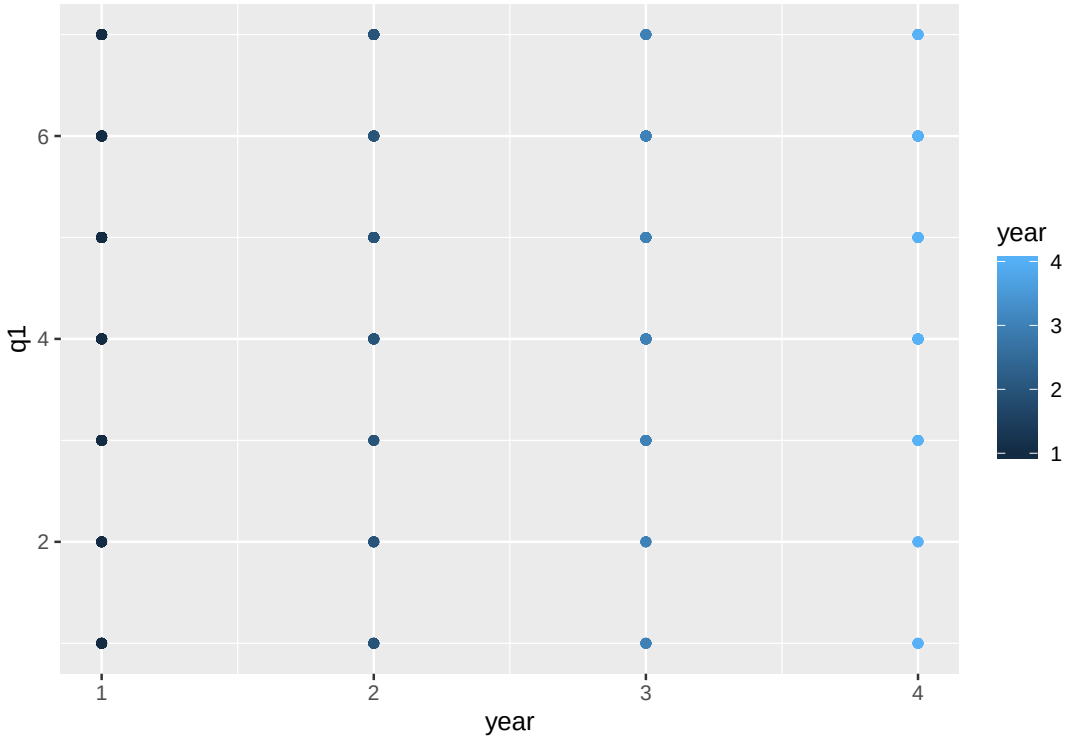

```
# relative importance by gender
ggplot(data = dat) +
  geom_bar(mapping = aes(x = mastgrad, fill = sex))
```




```
#subject by gender and normalizing using position = "dodge" to place overlapping objects directly
ggplot(data = dat) +
  geom_bar(mapping = aes(x = sex, fill = subject), position = "dodge")
```




```
# plot answers to q1 with relation to the student's year
ggplot(data = dat) +
  geom_point(mapping = aes(x = year, y = q1, colour = year))
```



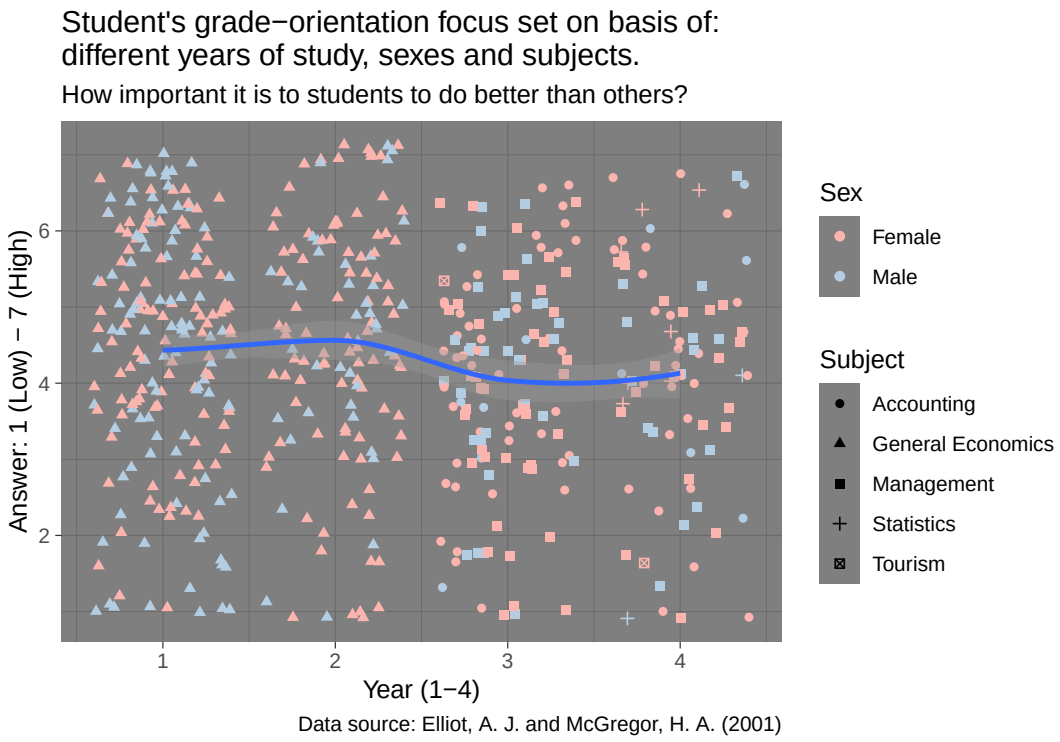

```
################################################################################
```

```
## Calculate mean for 7 categories:
# across 7 categories:
# - q1, q2, q3 - Performance approach questions
# - q4, q5, q6 - Performance avoidance questions
# - q7, q8, q9 - Mastery-Approach
# - q10, q11, q12 - Mastery-Avoidance
# - Interest
# - Enjoyment
# - Understanding/Grades

mean_dat <- dat
# get mean from q1, q2, q3 columns (Performance approach questions) for all the students
# save the results in 'm1' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m1 = pmap_dbl(select(., c("q1", "q2", "q3")), function(...) mean(c(...))))

# get mean from q4, q5, q6 columns (Performance avoidance questions) for all the students,
# save the results in 'm2' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m2 = pmap_dbl(select(., c("q4", "q5", "q6")), function(...) mean(c(...))))

# get mean from q7, q8, q9 columns (Mastery approach questions) for all the students
```

```r
# save the results in 'm3' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m3 = pmap_dbl(select(., c("q7", "q8", "q9")), function(...) mean(c(...))))

# get mean from q10, q11, q12 columns (Mastery avoidance questions) for all the students
# save the results in 'm4' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m4 = pmap_dbl(select(., c("q10", "q11", "q12")), function(...) mean(c(...))))

# get mean from 'interest' column (Course interestedness expectations) for all the students
# save the results in 'm_interest' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m_interest = pmap_dbl(select(., c("interest")), function(...) mean(c(...))))

# get mean from 'enjoy' column (Course enjoyment expectations) for all the students
# save the results in 'm_interest' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m_enjoy = pmap_dbl(select(., c("enjoy")), function(...) mean(c(...))))

# get mean from 'mastgrad' column (1 (Understanding) - 7 (Grades) Importance) for all the studen
# save the results in 'm_interest' colum and add it to 'mean_dat' table
mean_dat <- mean_dat %>%
  mutate(m_mastgrad = pmap_dbl(select(., c("mastgrad")), function(...) mean(c(...))))

# save final cleaned table
write_csv(mean_dat, "data/MeanCleanedStudentGoals.csv")
# save final cleaned table as tibble table
dat_tibble <- as_tibble(mean_dat)
```

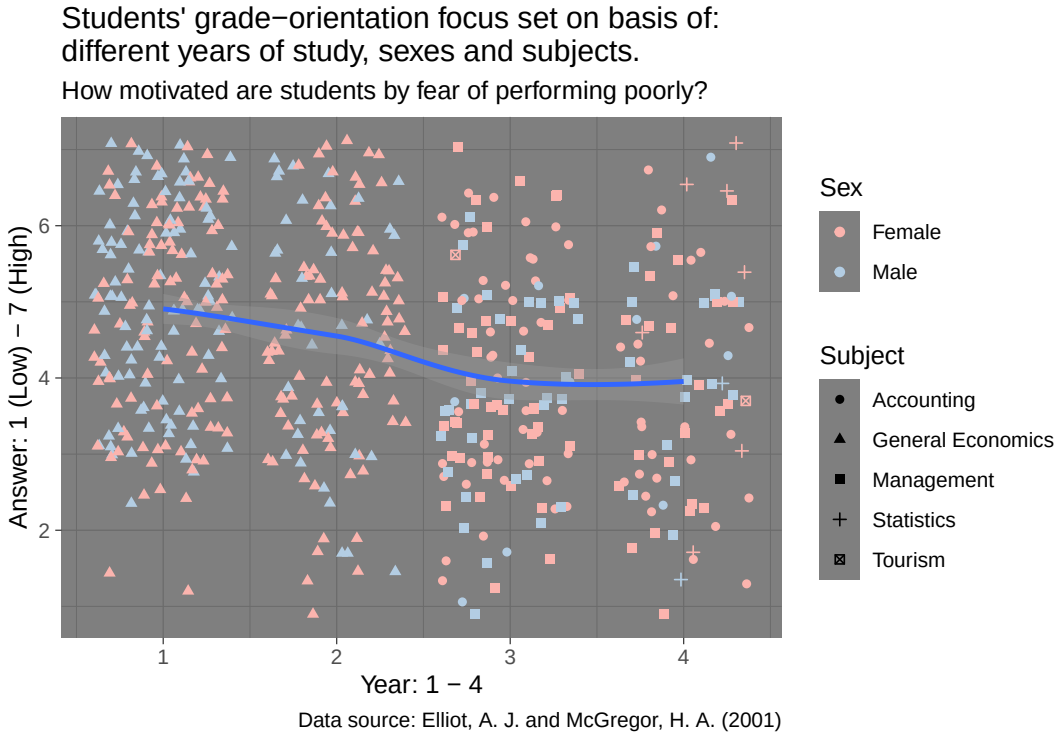
```r
# m1
# Plot mean results of performance approach questions
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(mean_dat$year, mean_dat$m1))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = loess, formula = y ~ x, se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
  title = "Student's grade-orientation focus set on basis of:
different years of study, sexes and subjects.",
  subtitle = "How important it is to students to do better than others?",
```

```
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year (1-4)",
  y = "Answer: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```



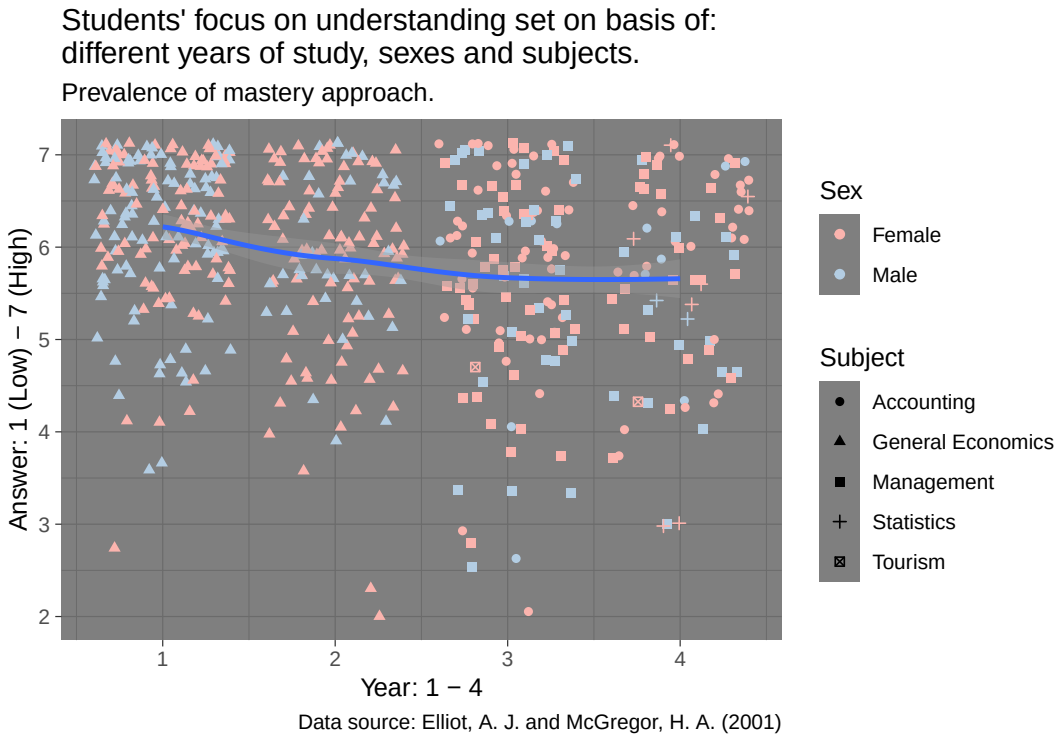

```
# m2
# Plot mean results of performance avoidance
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m2))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
  title = "Students' grade-orientation focus set on basis of:
```

```
different years of study, sexes and subjects.",
  subtitle = "How motivated are students by fear of performing poorly?",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year: 1 - 4",
  y = "Answer: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```



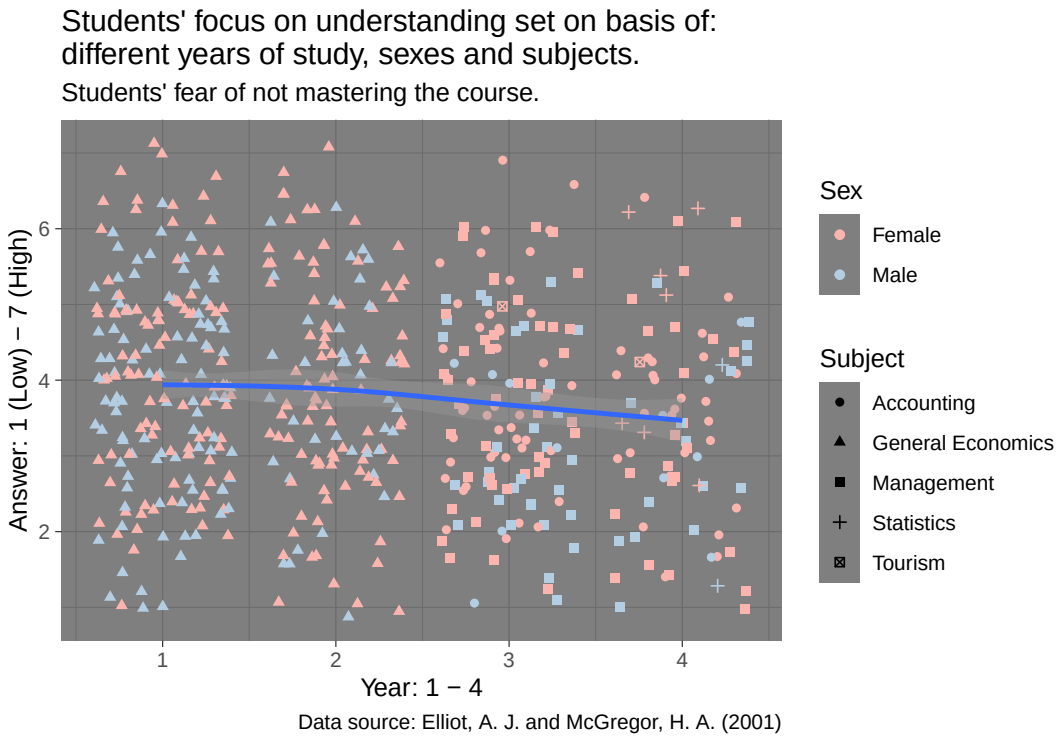

```
# m3
# Plot mean results of mastery approach questions
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m3))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = stats::loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
```

```
  title = "Students' focus on understanding set on basis of:
different years of study, sexes and subjects.",
  subtitle = "Prevalence of mastery approach.",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year: 1 - 4",
  y = "Answer: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```



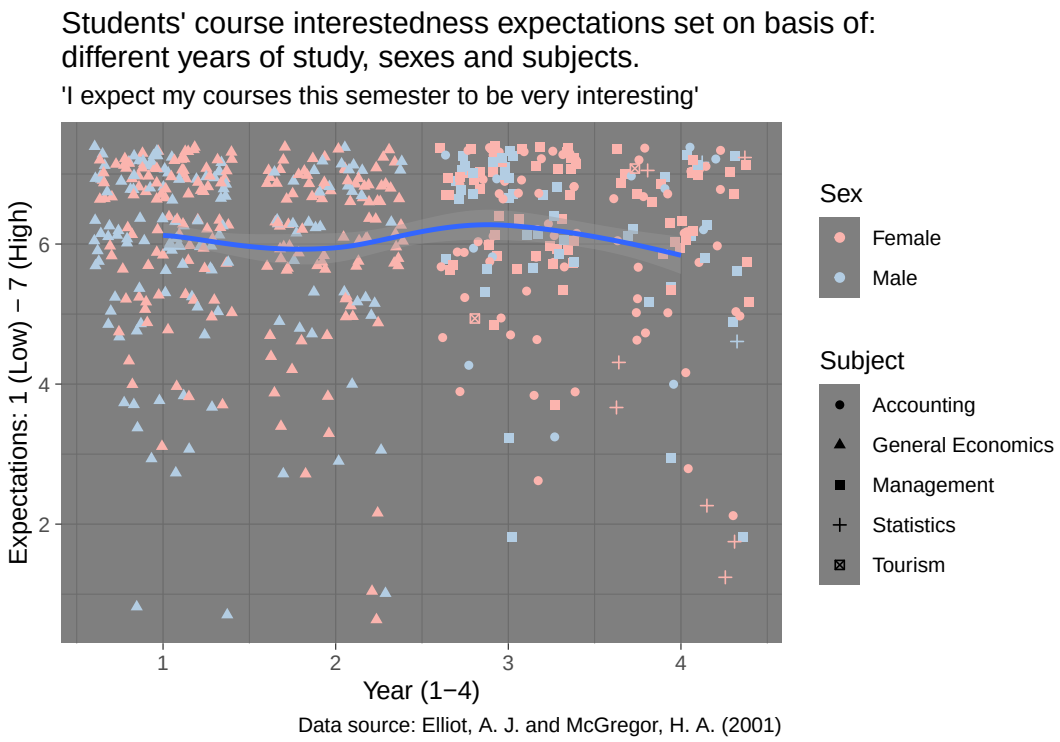

```
# m4
# Plot mean results of mastery avoidance questions
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m4))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = stats::loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
```

```
c + labs(
  title = "Students' focus on understanding set on basis of:
different years of study, sexes and subjects.",
  subtitle = "Students' fear of not mastering the course.",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year: 1 - 4",
  y = "Answer: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```

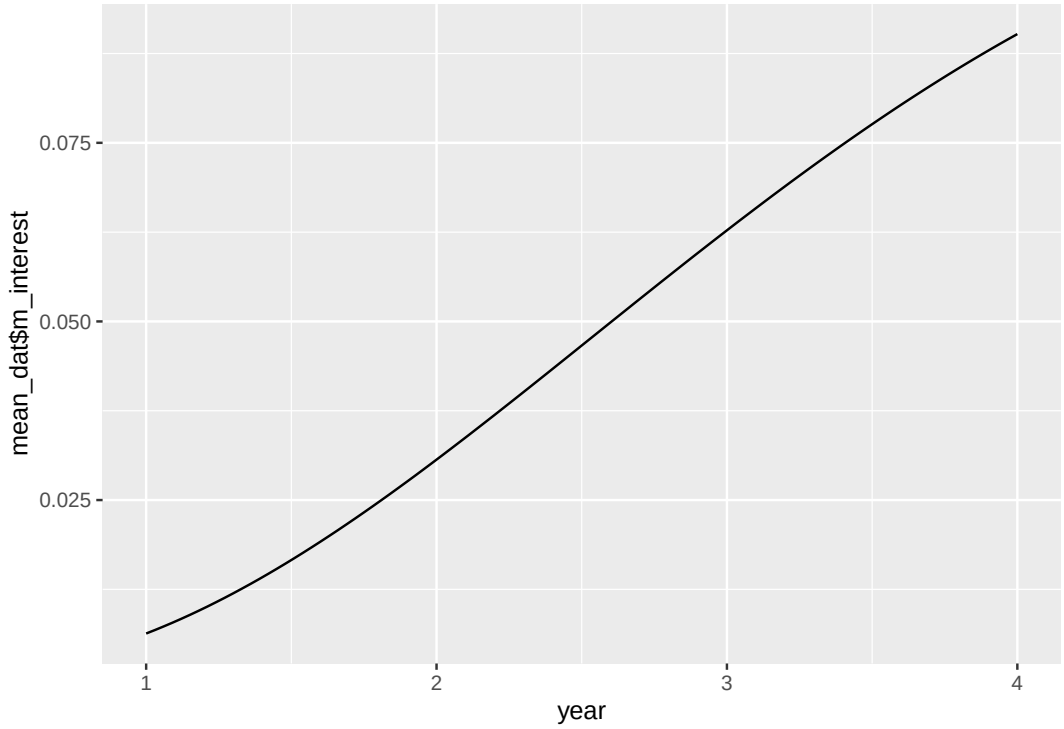



```
# interest
# Plot mean results of course interestedness expectations questions
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m_interest))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = stats::loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
```

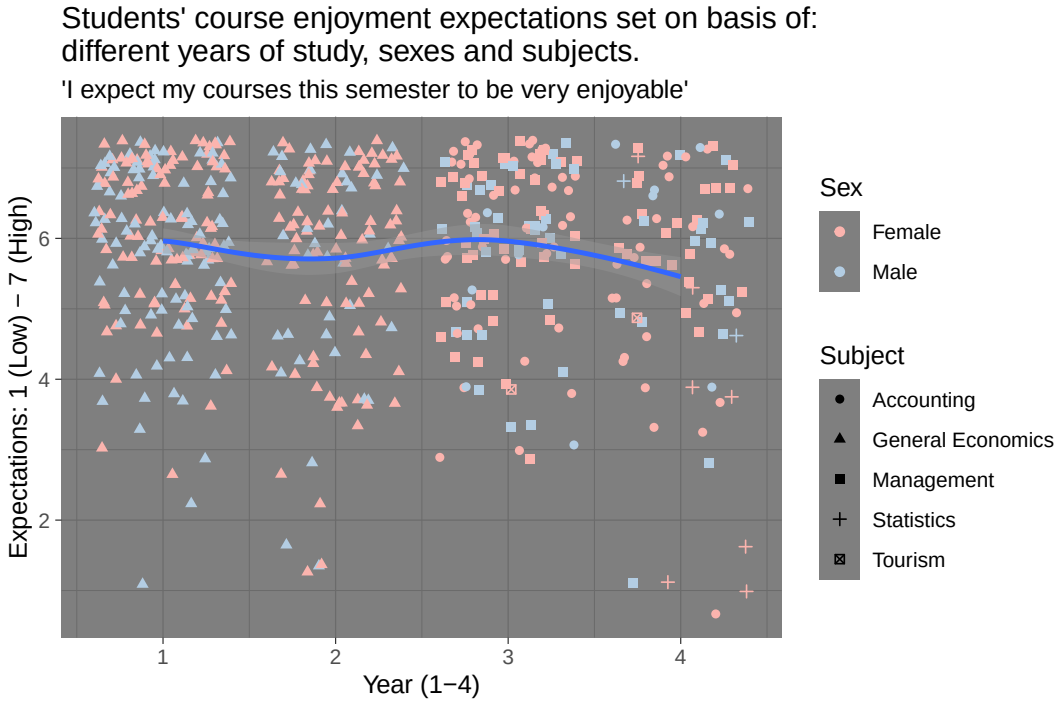
```
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
  title = "Students' course interestedness expectations set on basis of:
different years of study, sexes and subjects.",
  subtitle = "\'I expect my courses this semester to be very interesting\'",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year (1-4)",
  y = "Expectations: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```



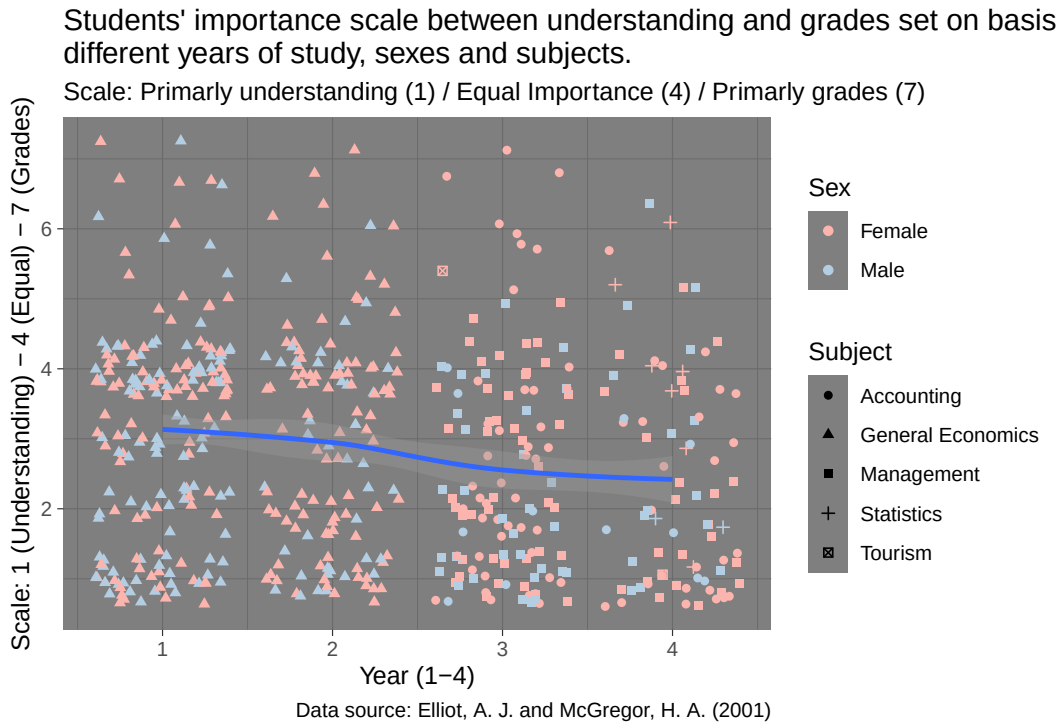

```
# chi-squared
chi_sqrt <- ggplot(data = dat, aes(year, mean_dat$m_interest)) +
  stat_function(fun = dchisq, args = list(df = 8))
chi_sqrt
```




```
# enjoy
# Plot mean results of course enjoyment expectations questions
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m_enjoy))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = stats::loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
  title = "Students' course enjoyment expectations set on basis of:
different years of study, sexes and subjects.",
  subtitle = "\'I expect my courses this semester to be very enjoyable\'",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year (1-4)",
  y = "Expectations: 1 (Low) - 7 (High)",
  colour = "Sex",
  shape = "Subject"
)
```




```
# mastgrad
# Plot mean results of (Primarly understanding/Equal Importance/Primarly grades)scale
# for all students with relation to student's year, sex and subject
# data
d <- ggplot(data = dat, aes(year, mean_dat$m_mastgrad))
# mapping data (use "jitter" to improve the graph and avoid gridding)
l <- d + geom_jitter(aes(colour = sex, shape = subject))
# smoothing
s <- l + geom_smooth(method = stats::loess, formula = y ~ log(x), se = TRUE)
# adding theme
t <- s + theme_dark()
# adding colouring
c <- t + scale_colour_brewer(palette = "Pastel1")
# adding labels
c + labs(
  title = "Students' importance scale between understanding and grades set on basis of:
different years of study, sexes and subjects.",
  subtitle = "Scale: Primarly understanding (1) / Equal Importance (4) / Primarly grades (7)",
  caption = "Data source: Elliot, A. J. and McGregor, H. A. (2001)",
  x = "Year (1-4)",
  y = "Scale: 1 (Understanding) - 4 (Equal) - 7 (Grades)",
  colour = "Sex",
  shape = "Subject"
)
```




```
## CLASSIFICATION ##########################################################
# dat_tibble %>%
#   head() %>%
#   knitr::kable()

# get only answers that are greater or equal to 5
dat_tibble_m1 <- filter(dat_tibble, m1 >= 5)

n_m1 <- tally(dat_tibble_m1) # 212
beta <- n_m1 / n # 0.3392

ci <- beta * ((1 - beta)/(n)) # 0.0003586294
ci_sqrt <- sqrt(ci) # 0.0189
ci_margin_error <- ci_sqrt * 1.96 # 0.0371 or 3.71%

# Our 95% confidence interval for the percentage of times we will get a student with a mean of
# 5 or above for the set of m1 questions is 0.3392 (or 34%), plus or minus 0.03711 (or 3.7%).
# The lower end of the interval is 0.3392 - 0.03711 which is:
lower_end_of_interval <- beta - ci_margin_error # 0.3020825 or 30%
# The upper end of the interval is 0.3392
upper_end_of_interval <- beta + ci_margin_error # 0.3763175 or 37%

# To interpret these results we could say that with 95% confidence the percentage of the times
# we should expect to find a student with a mean score of 5 or above to m1 is somewhere
```

```
# between 30% and 37%, based on our sample.
#####################################################################
```