Tellez de Ita

Rodrigo Tellez de Ita

```
here::i_am("grades.Rproj")
library(here)
library(dplyr)
library(readr)
library(vroom)
library(tidyr)
library(ggplot2)
library(stringr)
```

1. Introduction

Link: https://github.com/Rodrigotdi/grades.git

1.1 Study organisation

Question 1: Loading of course

```
course <- vroom(here("courses.csv"))
```

Question 2

```
course_table <- course|> rename ("course name"=course)|>rename(identifier=course_id)|> renam
course_table
```

course name	identifier	trimester	number of exams
Ancient Magic and Mysticism	1	3	3
Potion Brewing and Herbology	2	2	7
Celestial Navigation and Astronomy	3	3	10
Dragon Lore and Taming	4	2	7
Elemental Mastery and Control	5	2	7
Illusion and Enchantment	6	1	6
Necromancy and Spirit Summoning	7	1	4
Runecrafting and Glyphwork	8	1	8
Swordsmanship and Martial Arts	9	3	6
History of the Arcane	10	2	4

1.2 Students

Question 3

```
students <- vroom(here("students.csv"),
                  col_types = cols(
  id = col_integer(),
  group = col_factor(),
  birth_date = col_date(format = ""),
  sex = col_factor()
))

paste(
  "The birth_date column of the students data frame is of class",
  class(students$birth_date)
)
```

```
[1] "The birth_date column of the students data frame is of class Date"
```

1.3 Grades

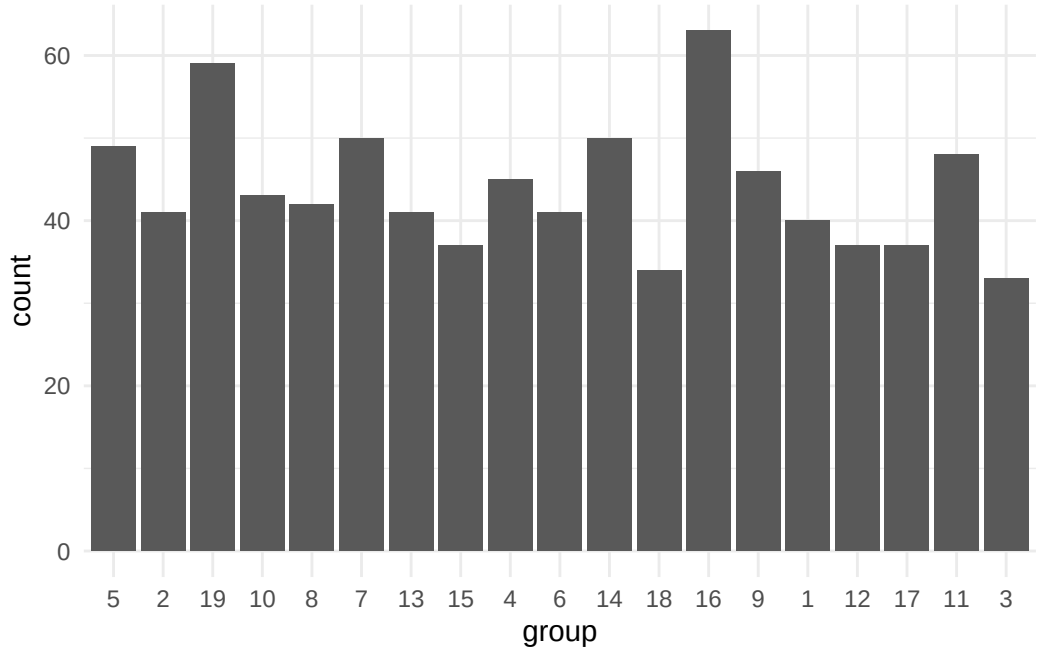
Question 4

```
grades <- read_delim(here("grades.csv"),delim=";",na="_")
```

2. Student population analysis

Question 5

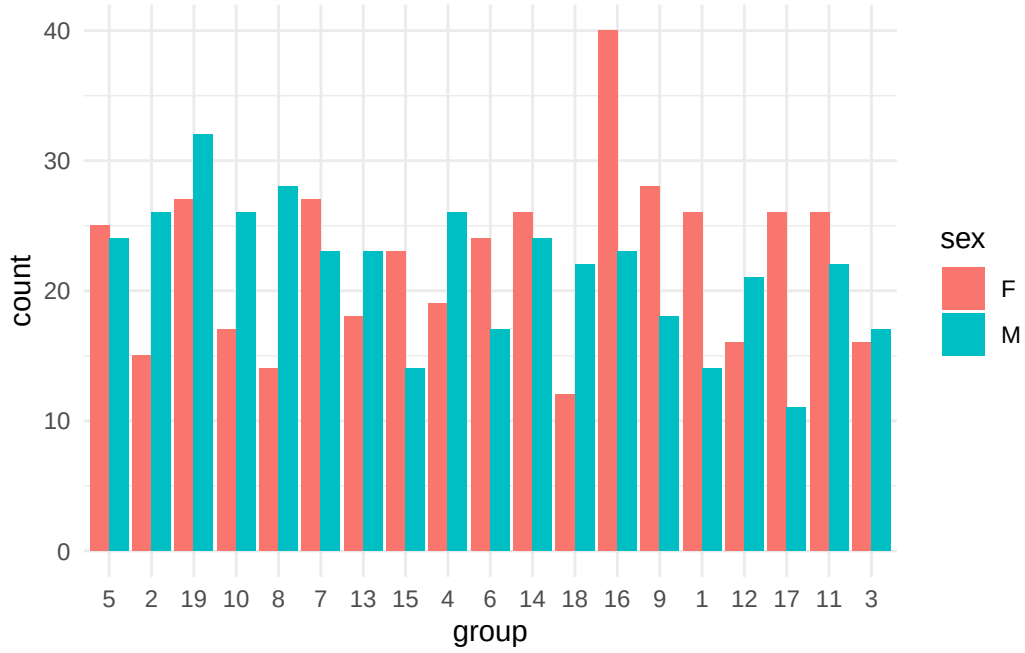
```
ggplot(students,aes(x=group))+
  geom_bar()+
  theme_minimal()
```



All the groups except 16 or 19 seems to have approximately the same number of students.

Question 6

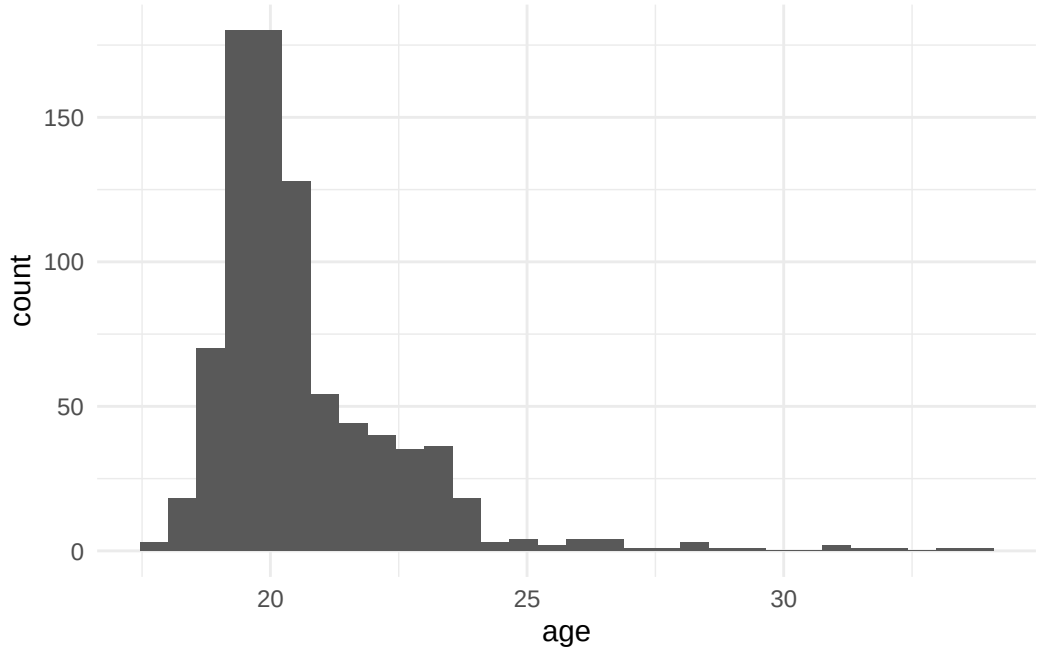
```
students |>
  ggplot(aes(x = group, fill = sex)) +
  geom_bar(position = "dodge")+
  theme_minimal()
```



For the majority of the group the gender are well balanced.

Question 7

```
library(lubridate)

students_age <- students |>
  mutate(age = time_length(today() - birth_date, unit = "year"))
ggplot(students_age, aes(x = age)) +
  geom_histogram() +
  theme_minimal()
```



Most of the students are around 20 years old.

Question 8

```
students_age |>
  summarise(
    median_age = round(median(age)),
    .by = group
  ) |>
  arrange(group) |> knitr::kable()
```

group	median_age
5	20
2	20
19	20
10	21
8	20
7	20
13	20
15	20

group	median_age
4	20
6	20
14	20
18	20
16	20
9	20
1	20
12	20
17	20
11	20
3	20

Question 9

```
oldest_per_group <- students_age |>
  slice_max(age, n = 1,by = group) |>
  mutate(age = round(age)) |>
  select(group,id, sex, age) |>
  arrange(desc(age))|>
  rename("oldest age"=age)|>
  knitr::kable()

oldest_per_group
```

group	id	sex	oldest age
1	279	F	34
8	411	M	32
16	701	M	32
7	114	M	31
9	199	F	31
6	75	F	29
12	223	M	29
17	741	M	28
14	153	M	26
11	190	M	26
3	613	F	26
13	539	M	25

group	id	sex	oldest age
15	329	M	25
4	209	M	25
18	296	M	25
5	210	M	24
2	174	M	24
10	718	M	24
19	80	M	23

3. Simple Grade analysis

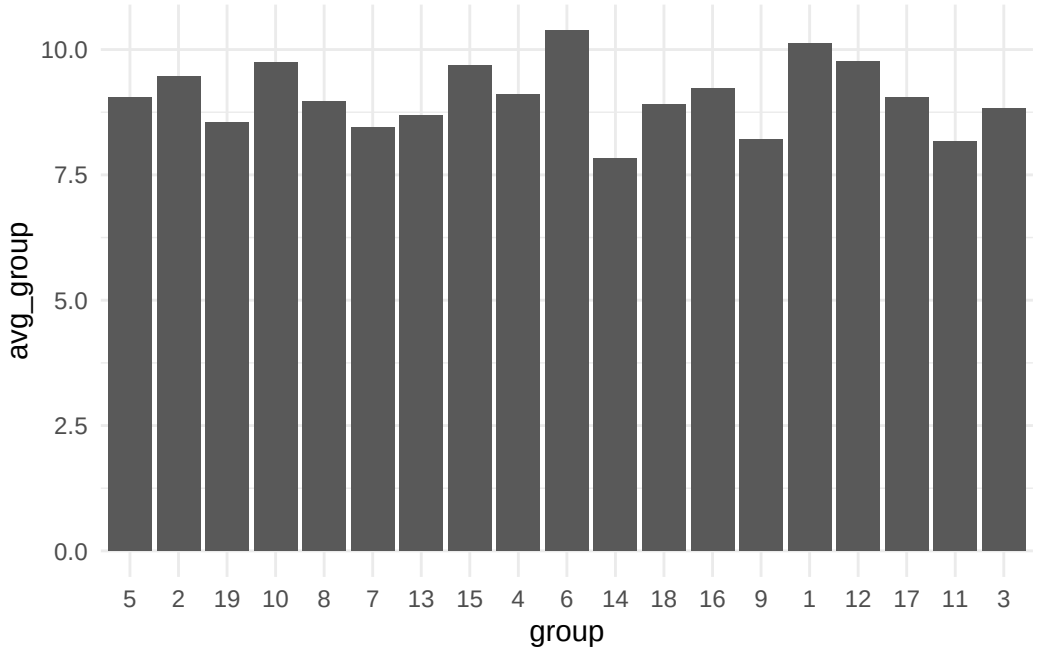
Question 10

```
num_grades <- grades |>
  filter(!is.na(grade)) |>
  nrow()
paste("The data set contains", num_grades, "grades.")
```

```
[1] "The data set contains 50264 grades."
```

Question 11

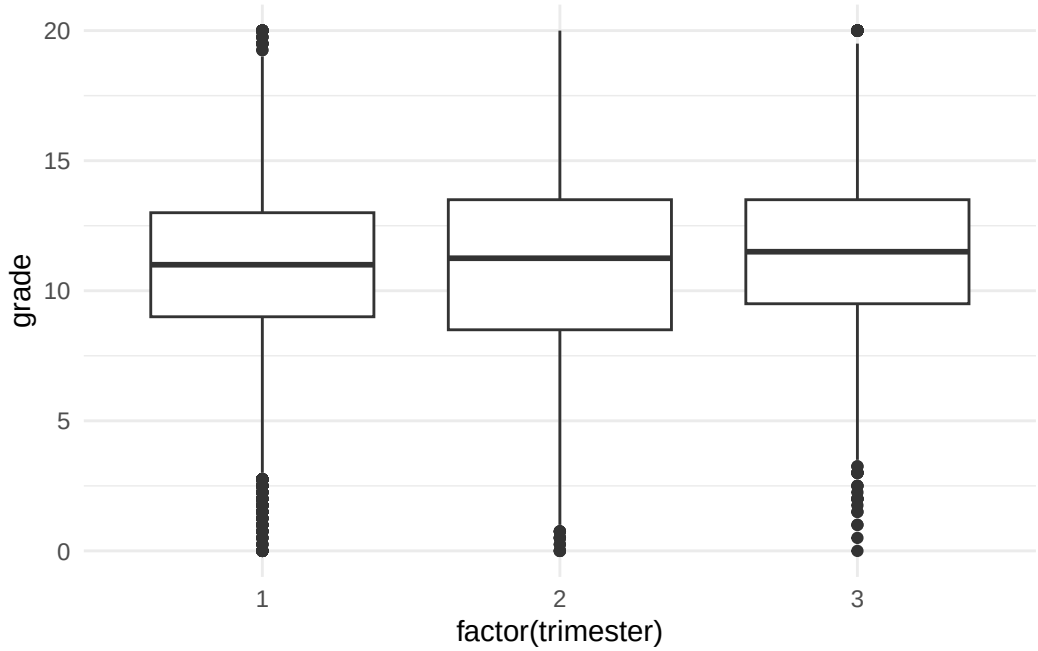
```
grades |>
  filter(!is.na(grade)) |>
  left_join(course |> select(course_id, course), by=join_by(course_id)) |>
  filter(course == "History of the Arcane") |>
  left_join(students |> select(id, group), by=join_by(id)) |>
  summarise(avg_group = mean(grade), .by = group) |>
  ggplot(aes(x = group, y = avg_group)) +
  geom_col()+
  theme_minimal()
```



All the groups have approximately the same averages in this course.

Question 12

```
grades |>
  filter(!is.na(grade)) |>
  left_join(
    course |> select(course_id, trimester), join_by(course_id)) |>
  ggplot(aes(x = factor(trimester), y = grade)) +
  geom_boxplot() +
  theme_minimal()
```



The grades seems to be the same around the different trimesters.

4. Attendance analysis

Question 13

```
grades_per_student <- grades |>
  filter(!is.na(grade)) |>
  summarise(num_grades = n(), .by = id) |>
  left_join(students |> select(id, group, sex), join_by(id))

grades_per_student|>
  slice_head(n = 5)
```

```
# A tibble: 5 x 4
     id num_grades group sex
  <dbl>      <int> <fct> <fct>
1    38         61 1     F
2    46         61 1     M
3    68         58 1     F
```

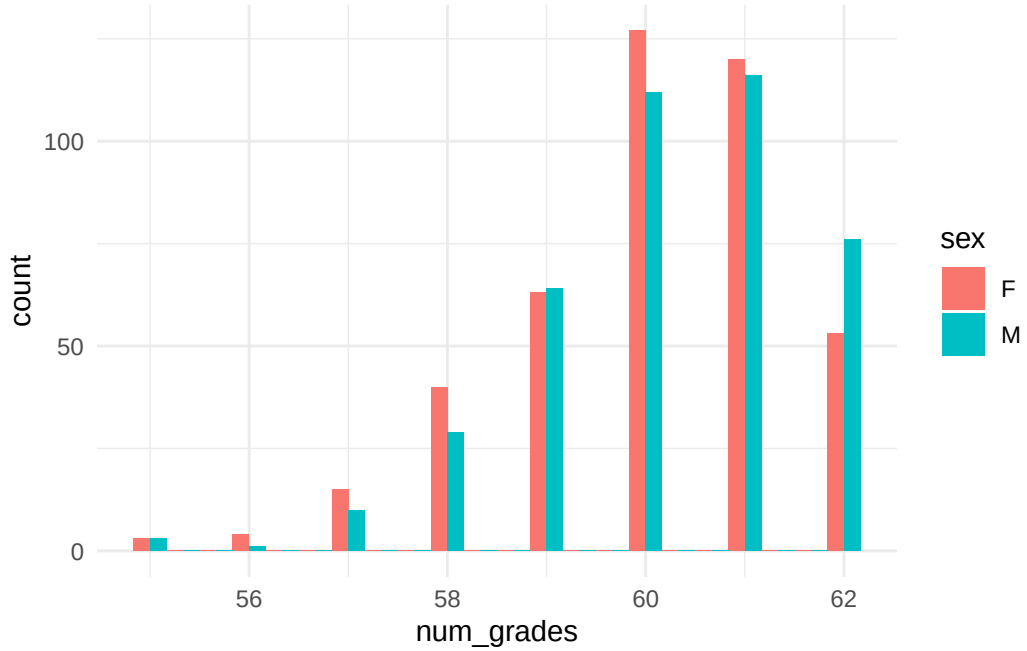
```
4   112          61 1     F
5   127          58 1     F
```

```
grades_per_student |>
  summarise(
    min_grades = min(num_grades),
    max_grades = max(num_grades),
    avg_grades = mean(num_grades),
    median_grades = median(num_grades)
  )|>knitr::kable()
```

min_grades	max_grades	avg_grades	median_grades
55	62	60.1244	60

Question 14

```
grades |>
  filter(!is.na(grade)) |>
  summarise(num_grades = n(), .by = id) |>
  left_join(students |> select(id, sex), join_by(id)) |>
  ggplot(aes(x = num_grades, fill = sex)) +
  geom_histogram(position = "dodge",bins=22) + theme_minimal()
```



There is no gender differences in the number of grades.

Question 15

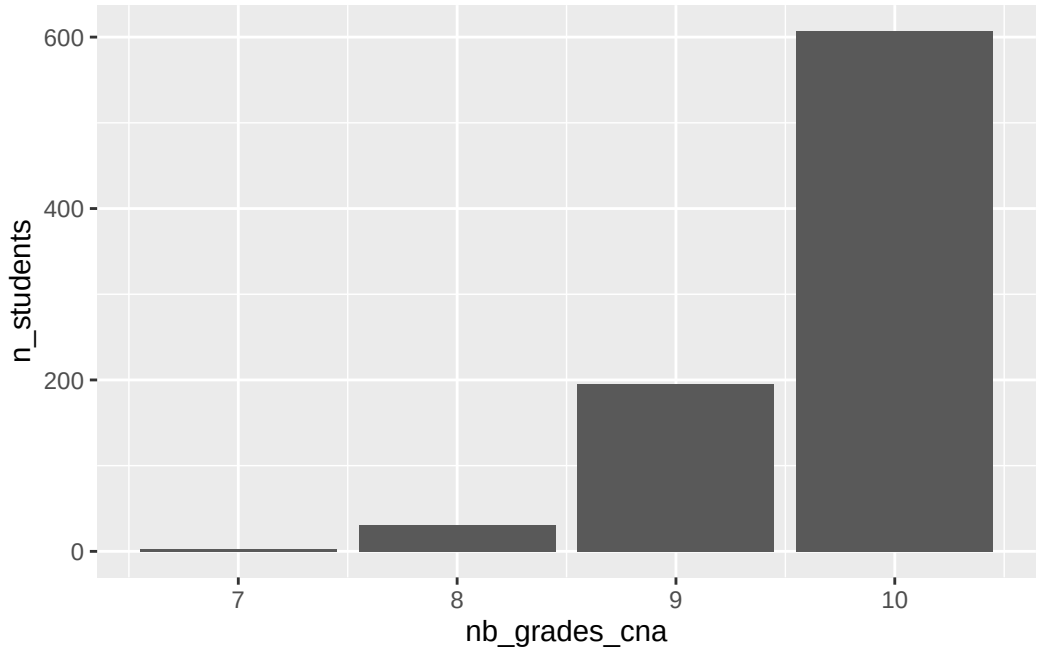
```
CNA <- grades |>
  filter(!is.na(grade)) |>
  left_join(course |> select(course_id, course),join_by(course_id)) |>
  filter(course == "Celestial Navigation and Astronomy") |>
  summarise(
    nb_grades_cna = n(),
    .by = id
  ) |>
  left_join(
    students |> select(id, group),
    join_by(id)
  ) |>
  select(id, group, nb_grades_cna)

CNA |>slice_head(n=5)
```

```
# A tibble: 5 x 3
```

```
     id group nb_grades_cna
  <dbl> <fct>          <int>
1    38 1                 10
2    46 1                 10
3    68 1                 10
4   112 1                 10
5   127 1                 10
```

Question 16

```
CNA|>summarise(n_students=n(), .by=nb_grades_cna) |> ggplot(aes(x=nb_grades_cna,y=n_students)
  geom_col()
```

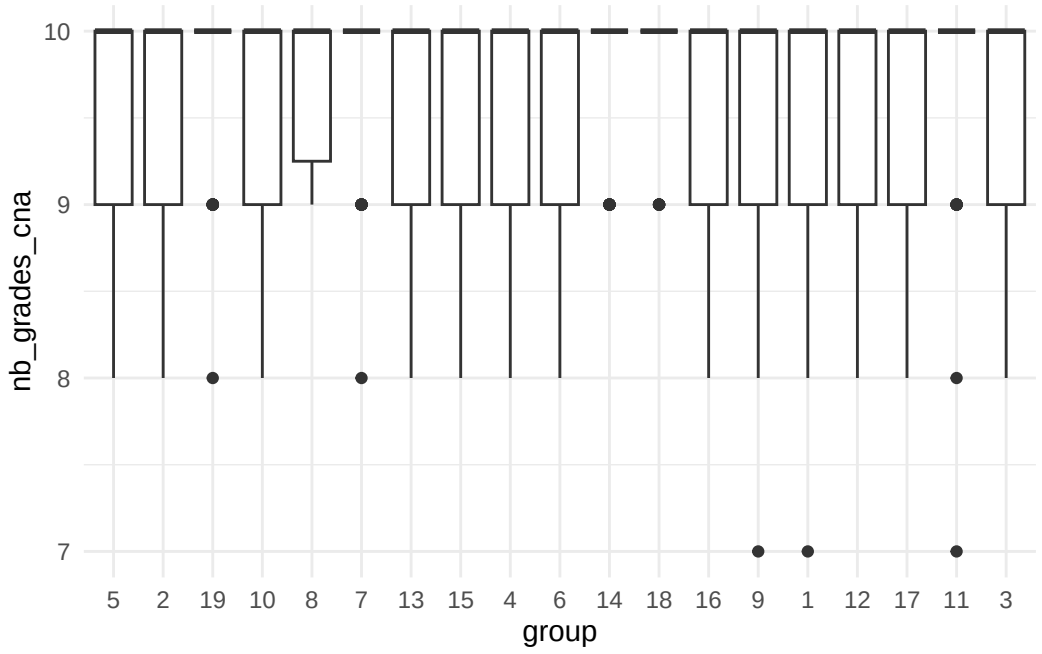


Most of the students have 10 grades.

Question 17

```
grades |>
  filter(!is.na(grade)) |>
```

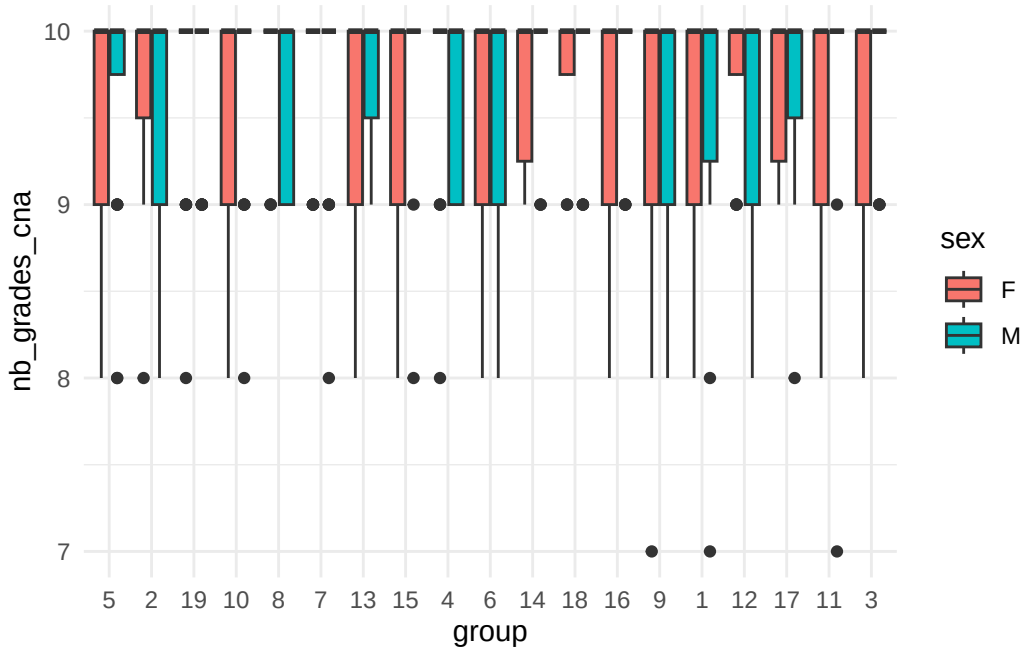
```
  left_join(
    course |> select(course_id, course),
    join_by(course_id)
  ) |>
  filter(course == "Celestial Navigation and Astronomy") |>
  summarise(
    nb_grades_cna = n(),
    .by = id
  ) |>
  left_join(
    students |> select(id, group),
    join_by(id)
  ) |>
  ggplot(aes(x = group, y = nb_grades_cna)) +
  geom_boxplot() + theme_minimal()
```



Question 18

```
grades |>
  filter(!is.na(grade)) |>
```

```r
left_join(
  course|> select(course_id, course),
  join_by(course_id)
) |>
filter(course == "Celestial Navigation and Astronomy") |>
summarise(
  nb_grades_cna = n(),
  .by = id
) |>
left_join(
  students |> select(id, group, sex),
  join_by(id)
) |>
ggplot(aes(x = group, y = nb_grades_cna, fill = sex)) +
geom_boxplot() + theme_minimal()
```



5. Grade analysis

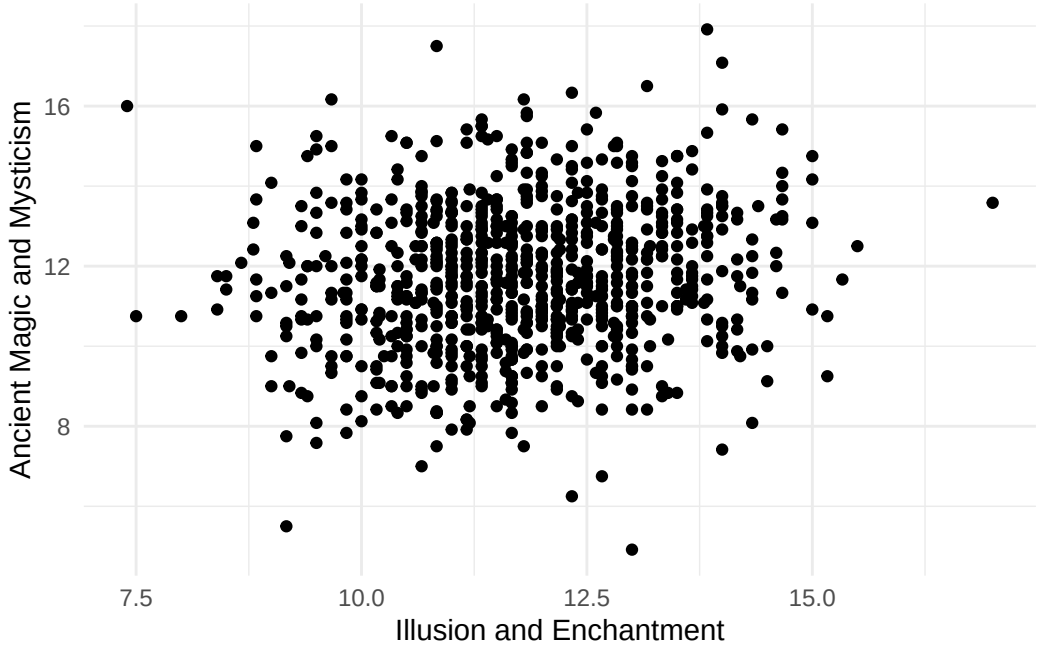
Question 19

```
avg_per_course <- grades |>
  filter(!is.na(grade)) |>
  left_join(
    course |> select(course_id, course),
    join_by(course_id)
  ) |>
  summarise(
    avg_grade = mean(grade),
    .by = c(id, course)
  ) |>
  left_join(
    students |> select(id, group),
    join_by(id)
  )


avg_wider <- avg_per_course |>
  select(id, group, course, avg_grade) |>
  pivot_wider(
    names_from = course,
    values_from = avg_grade
  )

avg_wider|>select(id,group,`Ancient Magic and Mysticism`,`Celestial Navigation and Astronomy
  slice_head(n=5)|>knitr::kable()
```

id	group	Ancient Magic and Mysticism	Celestial Navigation and Astronomy
38	1	13.66667	13.80
46	1	12.91667	13.95
68	1	14.83333	12.80
112	1	14.33333	12.70
127	1	10.87500	13.10

Question 20

```
avg_wider|> ggplot(aes(x=`Illusion and Enchantment`,y=`Ancient Magic and Mysticism`))+
  geom_point()+theme_minimal()
```



The correlation between both classes is not very clear. Some of the students have very good grades in both courses but the grade in one course does seem to be close on the other course for everyone, at least on this graph. Let's check the correlations.

Question 21

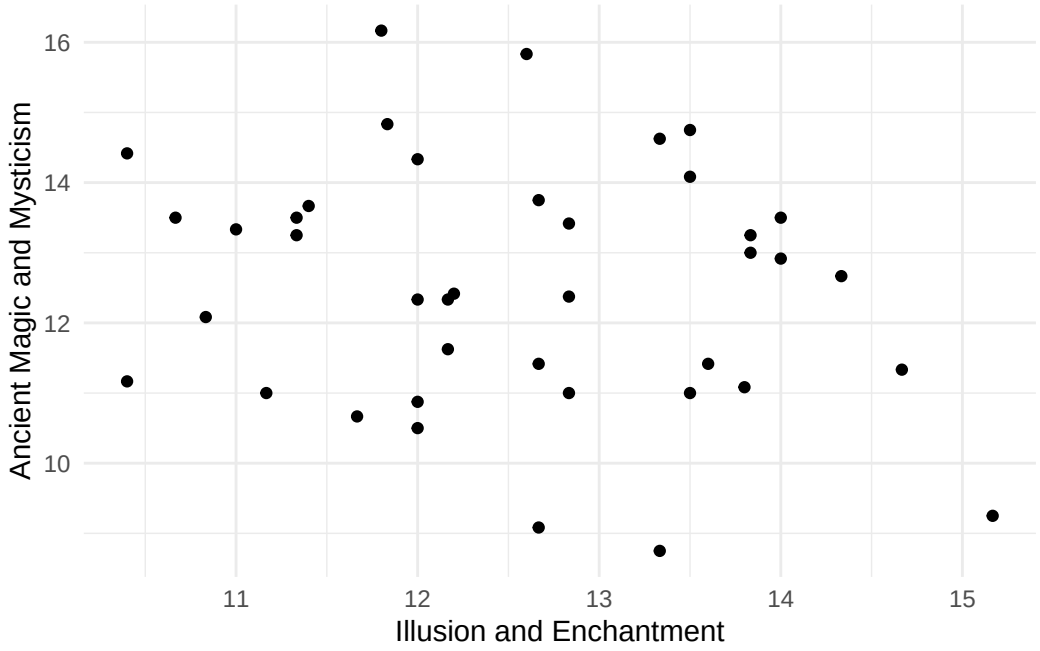
```
avg_wider |> summarise(cor_ill_cna=cor(`Illusion and Enchantment`,`Celestial Navigation and
```

group	cor_ill_cna
1	0.7432685
2	0.6539226
3	0.4759441
4	0.5758219
5	0.5851078

group	cor_ill_cna
6	0.5891706
7	0.3523564
8	0.5351540
9	0.6099001
10	0.6168292
11	0.4115246
12	0.6967605
13	0.3168674
14	0.3711671
15	0.6063386
16	0.4794431
17	0.3188879
18	0.4739535
19	0.5638979

Question 22

```
most_corr_group <- avg_wider |> summarise(cor_ill_cna=cor(`Illusion and Enchantment`,`Celesti

avg_wider|>semi_join(most_corr_group,by=join_by(group))|>ggplot(aes(x=`Illusion and Enchantme
  geom_point()+theme_minimal()
```



Even in the most correlated group it's hard to see a true correlation between both courses.

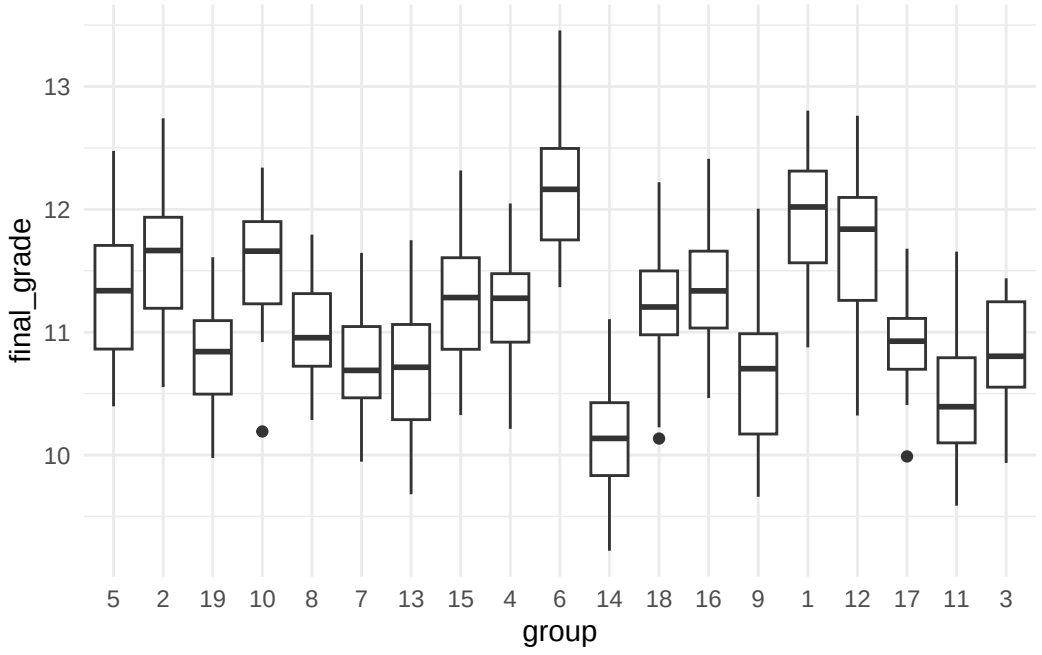
Question 23

```
final_grades <- grades |>
  filter(!is.na(grade)) |>
  summarise(
    avg_course = mean(grade),
    .by = c(id, course_id)
  ) |>
 summarise(
    final_grade = mean(avg_course),
    .by = id
  ) |> left_join(
    students |> select(id, group, sex),
    join_by(id)
  ) |>
  select(id, group, sex, final_grade) |>
  arrange(desc(final_grade))

final_grades|>slice_head(n=5)|>knitr::kable()
```

id	group	sex	final_grade
622	6	M	13.45625
479	6	F	13.41152
653	6	M	13.40509
67	6	F	13.29479
785	6	F	13.14863

Question 24

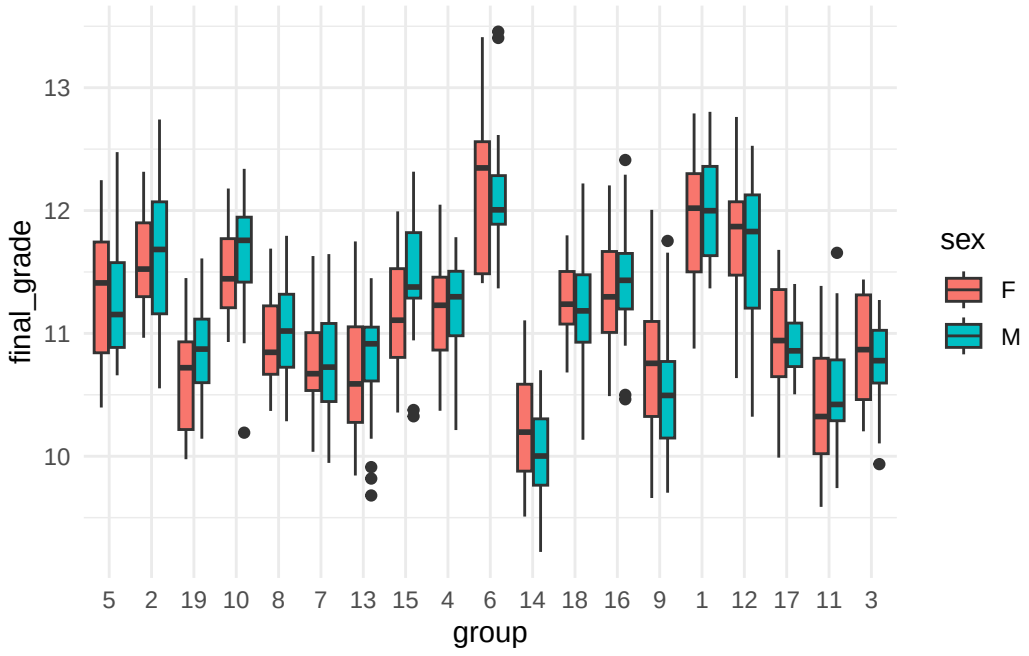
```
final_grades |>
  ggplot(aes(x = group, y = final_grade)) +
  geom_boxplot() + theme_minimal()
```



Most of the groups have similar distributions of final grades except for some group that are very week (14) or strong (6) so maybe groups can have an impact on having good grades.

Question 25

```r
final_grades |>
  ggplot(aes(x = group, y = final_grade, fill = sex)) +
  geom_boxplot() + theme_minimal()
```



Inside each group the gender doesn't seem to influence the final grades.

Question 26

```r
per_course_avg <- grades |>
  filter(!is.na(grade)) |>
  summarise(
    avg_course = mean(grade),
    .by = c(id, course_id)
  ) |>
  left_join(
    course |> select(course_id, trimester),
    join_by(course_id)
  )
```

```r
pass_cond1 <- per_course_avg |>
  summarise(
    final_grade    = mean(avg_course),
    min_course_avg = min(avg_course),
    .by = id
  )


trimester_cond <- per_course_avg |>
  summarise(
    trim_avg = mean(avg_course),
    .by = c(id, trimester)
  ) |>
  summarise(
    min_trim_avg = min(trim_avg),
    .by = id
  )
```

```r
pass_df <- pass_cond1|>left_join(trimester_cond,by=join_by(id))|>mutate(
  pass=(min_course_avg>=5) & (min_trim_avg>=10))|>
  left_join(students,by=join_by(id))|>select(id,group,final_grade,pass)

pass_df|>slice_head(n=10)|>knitr::kable()
```

id	group	final_grade	pass
38	1	12.30780	TRUE
46	1	12.27952	TRUE
68	1	11.99458	TRUE
112	1	12.32208	TRUE
127	1	11.50464	TRUE
129	1	11.60610	TRUE
192	1	12.43628	TRUE
202	1	12.79051	TRUE
229	1	12.04402	TRUE
231	1	12.64330	TRUE

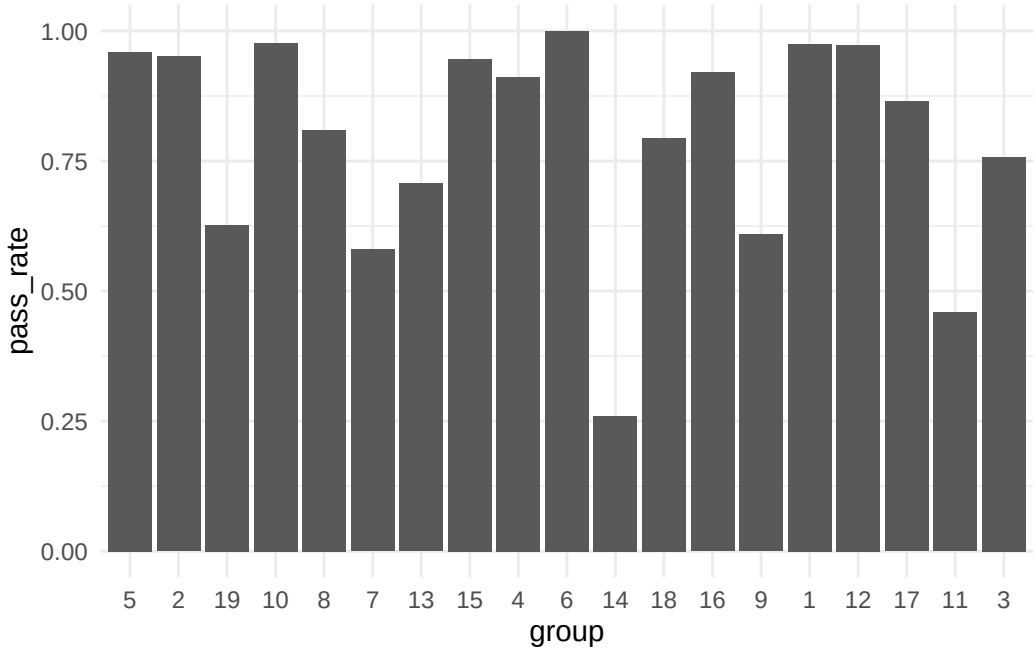
Question 27

```r
nb_fail <- pass_df|>filter(!pass,final_grade>=10)|>nrow()

paste(
  "There are",
  nb_fail,
  "students who do not pass but have a final grade greater or equal to 10."
)
```

```
[1] "There are 137 students who do not pass but have a final grade greater or equal to 10."
```

Question 28

```r
pass_rate <- pass_df|>summarise(pass_rate=mean(pass),.by=group)

pass_rate|> ggplot(aes(x=group,y=pass_rate))+geom_col()+theme_minimal()
```



The passing rate is high for every group except 14 and 11 that are below 50%.

Question 29

```
per_course <- grades |>
  filter(!is.na(grade)) |>
  summarise(
    avg_course = mean(grade),
    .by = c(id, course_id)
  ) |>
  left_join(
    course |> select(course_id, course),
    join_by(course_id)
  )

course_fail_counts <- per_course |>
  filter(avg_course < 5) |>
  summarise(
    n_students = n(),
    .by = course
  ) |>
  arrange(desc(n_students))

course_fail_counts|> knitr::kable()
```

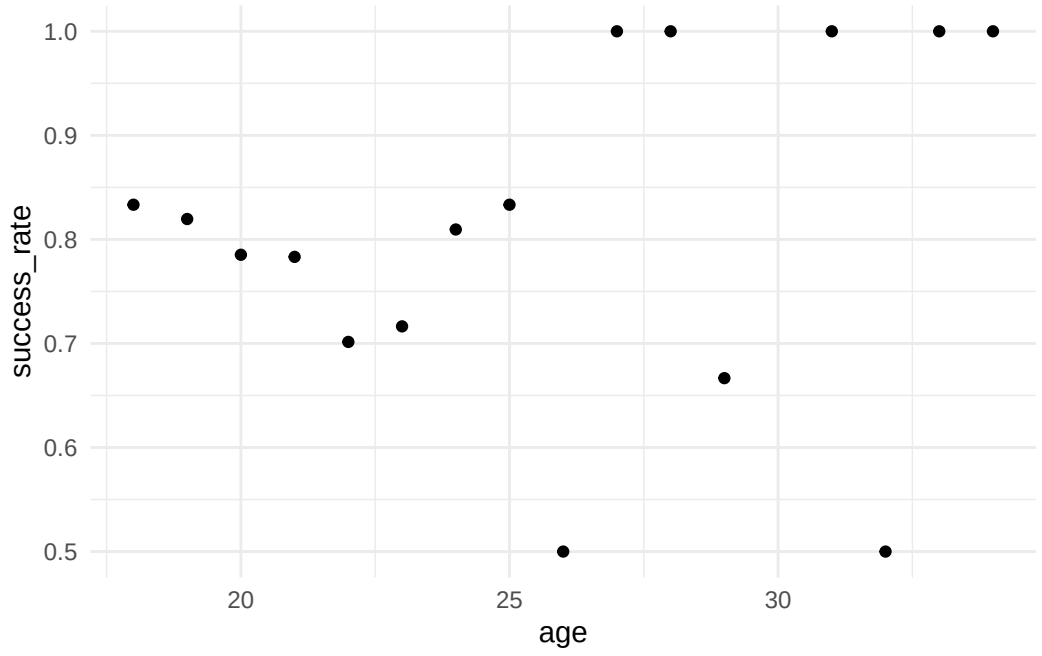
course	n_students
History of the Arcane	2
Ancient Magic and Mysticism	1

History of the Arcane and Ancient Magic and Mysticism are the most difficult courses, even if the number of students below 5 is very low.

Question 30

```
pass_age <- pass_df |>
  left_join(
    students |> select(id, birth_date),
    join_by(id)
  ) |>
  mutate(
```

```
    age = time_length(today() - birth_date, unit = "year"),
    age = round(age)
  )


success_by_age <- pass_age |>
  summarise(
    success_rate = mean(pass),
    .by = age
  ) |>
  arrange(age)


success_by_age |>
  ggplot(aes(x = age, y = success_rate))+
  geom_point() + theme_minimal()
```



The age does not really influence the success rate. In addition most of the students are around 20 years old so there are not enough "old" students to really see if age impacts success rate.