Main

Project 2

```
library(tidyverse)
library(tidymodels)

students <- read_delim("students.csv", delim = ";")

glimpse(students)
```

```
Rows: 649
Columns: 33
$ 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        <dbl> 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       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
$ Fedu       <dbl> 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 <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
$ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
$ failures   <dbl> 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     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
$ freetime   <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,~
$ goout      <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,~
$ Dalc       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
$ Walc       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,~
$ health     <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,~
$ absences   <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2, 2, 6~
$ G1         <dbl> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12, 14, 1~
$ G2         <dbl> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, 12, 14,~
$ G3         <dbl> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13, 15,~
```

Here we form 2 groups according to their level of alcohol assumption

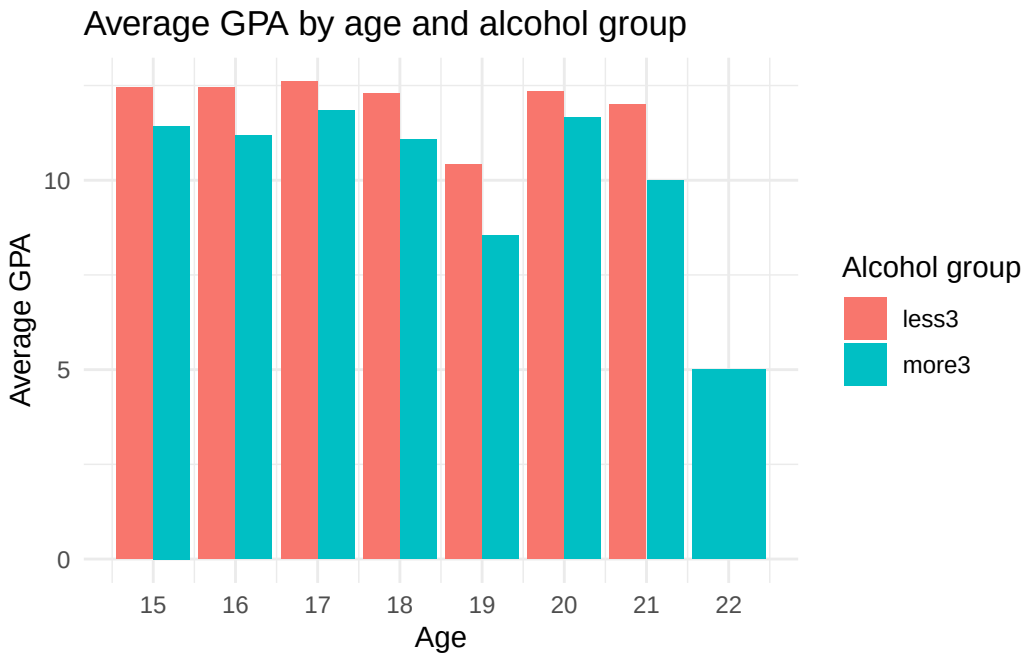
```
group1 <- students %>% filter(Dalc < 3 & Walc < 3)
group2 <- students %>% filter(Dalc >= 3 | Walc >= 3)

cat("Number of students with less than rate 3 of alcohol consumption:", nrow(group1), "\n")
```

```
Number of students with less than rate 3 of alcohol consumption: 391
```

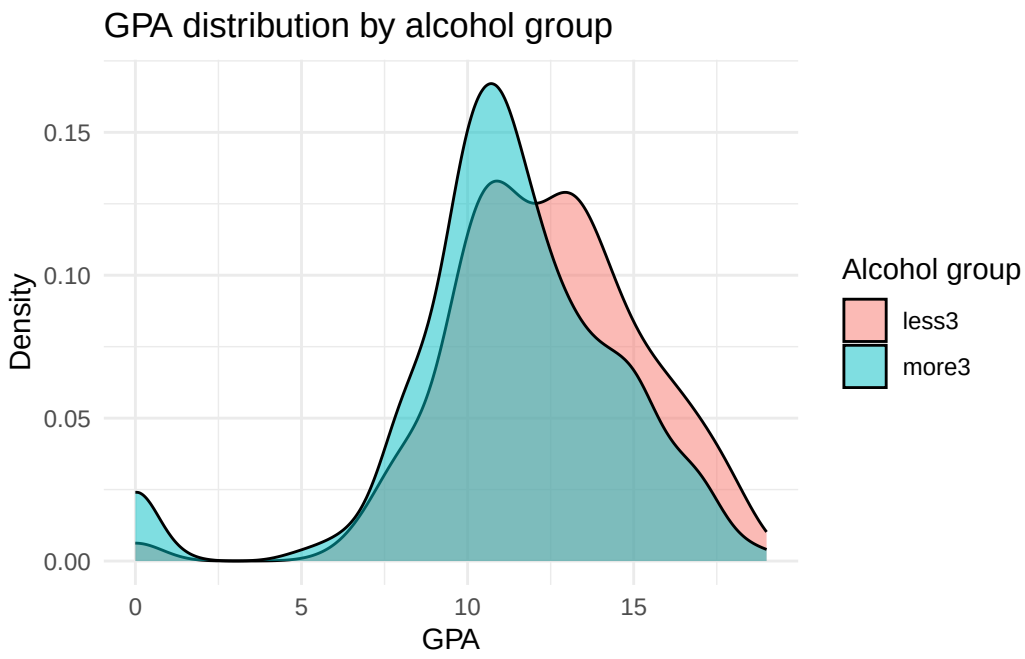
```
cat("Number of students with greater than rate 3 of alcohol consumption:", nrow(group2), "\n"
```

```
Number of students with greater than rate 3 of alcohol consumption: 258
```

```
students <- students %>%
  mutate(alcohol_group = ifelse(Dalc < 3 & Walc < 3, "less3", "more3"))
```




```
ggplot(students, aes(x=G3,fill=alcohol_group))+
  geom_density(alpha=0.5)+
  labs(title="GPA distribution by alcohol group",
       x="GPA",
       y="Density",
       fill="Alcohol group")+
  theme_minimal()
```




```
statistics <- students %>%
  group_by(alcohol_group) %>%
  summarize(
    mean_delivery_time = mean(G3),
    median_delivery_time = median(G3),
    sd_delivery_time = sd(G3),
    min_delivery_time = min(G3),
    max_delivery_time = max(G3),
    count = n()
  )
statistics
```

```
# A tibble: 2 x 7
  alcohol_group mean_delivery_time median_delivery_time sd_delivery_time
  <chr>                      <dbl>                <dbl>            <dbl>
1 less3                       12.4                   12             3.00
2 more3                       11.2                   11             3.45
# i 3 more variables: min_delivery_time <dbl>, max_delivery_time <dbl>,
#   count <int>
```

We see that

```r
control_count <- statistics$count[1]
test_count <- statistics$count[2]

count_difference <- abs(control_count-test_count)

cat("Count of less than 3 group:", control_count, "\nCount of more than 3 group:", test_count
```

```
Count of less than 3 group: 391
Count of more than 3 group: 258
Difference of counts: 133
```

. . .

```r
control_mean <- statistics$mean_delivery_time[1]
test_mean <- statistics$mean_delivery_time[2]
cat("Mean of less than 3 group:", control_mean, "\nMean of more than 3 group:", test_mean)
```

```
Mean of less than 3 group: 12.36573
Mean of more than 3 group: 11.2093
```

. . .

```r
control_sd <- statistics$sd_delivery_time[1]
test_sd <- statistics$sd_delivery_time[2]
cat("Standard deviation of less than 3 group:",control_sd, "\nStandard deviation of more than
```

```
Standard deviation of less than 3 group: 2.997735
Standard deviation of more than 3 group: 3.445347
```

. . .

```r
t.test(G3 ~ alcohol_group, data = students)
```

```
    Welch Two Sample t-test

data:  G3 by alcohol_group
t = 4.4027, df = 496.28, p-value = 1.31e-05
alternative hypothesis: true difference in means between group less3 and group more3 is not e
```

```
95 percent confidence interval:
 0.6403555 1.6724977
sample estimates:
mean in group less3 mean in group more3
           12.36573            11.20930
```

The p-value is much smaller than 0.05, which means that we can reject the null hypothesis. The conclusion is that the GPA for the groups is different. In our assumption it means that the GPA of those who consume more alcohol is less than of those who consume less alcohol.

We can also see that the confidence interval for the difference in means is ( ). Since zero value is not within this interval, we can reject the null hypothesis.

```r
get_stat_power <- function(N1,N2, mu_h0, mu_alternative,sd_h0, sd_alternative, alpha) {

  norm_h0 <- dnorm(0:N1, mean = mu_h0, sd = sd_h0)
  norm_alternative <- dnorm(0:N2, mean = mu_alternative, sd = sd_alternative)

  critical_value <- qnorm(1 - alpha, mean = mu_h0, sd = sd_h0) + 1

  power <- 1 - pnorm(critical_value - 1, mean = mu_h0, sd = sd_alternative)

  return(power)
}

stat_power <- get_stat_power(control_count, test_count, control_mean, test_mean, control_sd,
cat("Statistical power:",stat_power)
```

```
Statistical power: 0.07619255
```

The statistical power is really small, which means that we can easily fail to reject the null hypothesis even if the alternative hypothesis is true (meaning there is a real difference in GPA between the less3 and more3 groups).