Portfolio Exercise 4

Zhongxuan Sun

5/8/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

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

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggsci)
library(patchwork)
```

Link to the original data website: https://www.kaggle.com/janiobachmann/math-students?select=student-mat.csv

My download link: https://uwmadison.box.com/shared/static/yyanm3kyol042p90zm8trhhzj1h7tvs3.csv

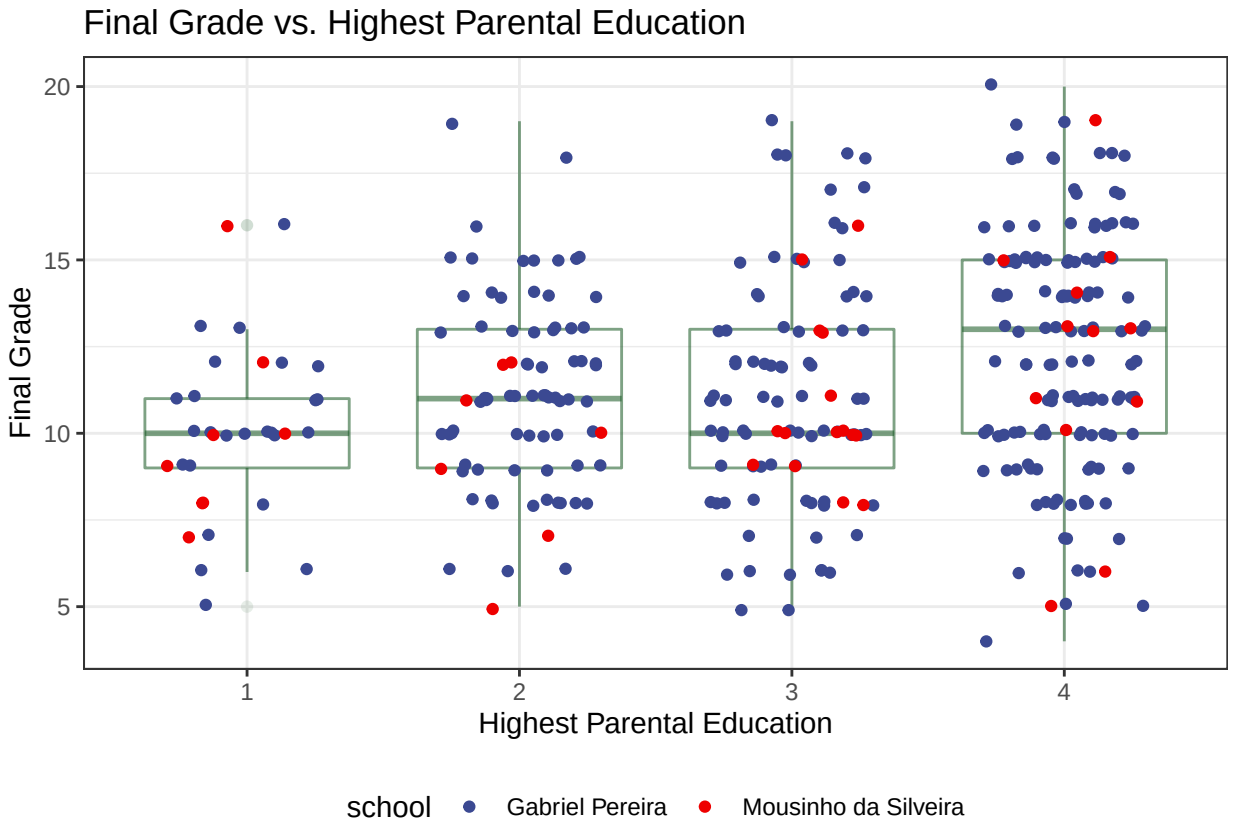
citation of the data:

```
df <- fread("https://uwmadison.box.com/shared/static/yyanm3kyol042p90zm8trhhzj1h7tvs3.csv")
df <- df %>% filter(G3 != 0)
```

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture Business TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Plot 1

```
set.seed(05082022)
df_points <- df %>%
  filter(G3 != 0) %>% # remove
  mutate(
    eduParent = factor(pmax(Medu, Fedu)),
    school = ifelse(school == "GP",
                    "Gabriel Pereira",
                    "Mousinho da Silveira")
  )
new_p1 <- df_points %>% ggplot(aes(x = eduParent, y = G3)) +
  geom_boxplot(alpha = 0.1, col = "#195a258f") +
  geom_jitter(aes(col = school), width = 0.3, height = 0.1) +
  theme_bw() +
  labs(x = "Highest Parental Education",
       y = "Final Grade",
       title = "Final Grade vs. Highest Parental Education") +
  theme(legend.position = "bottom") +
  scale_color_aaas()
new_p1
```

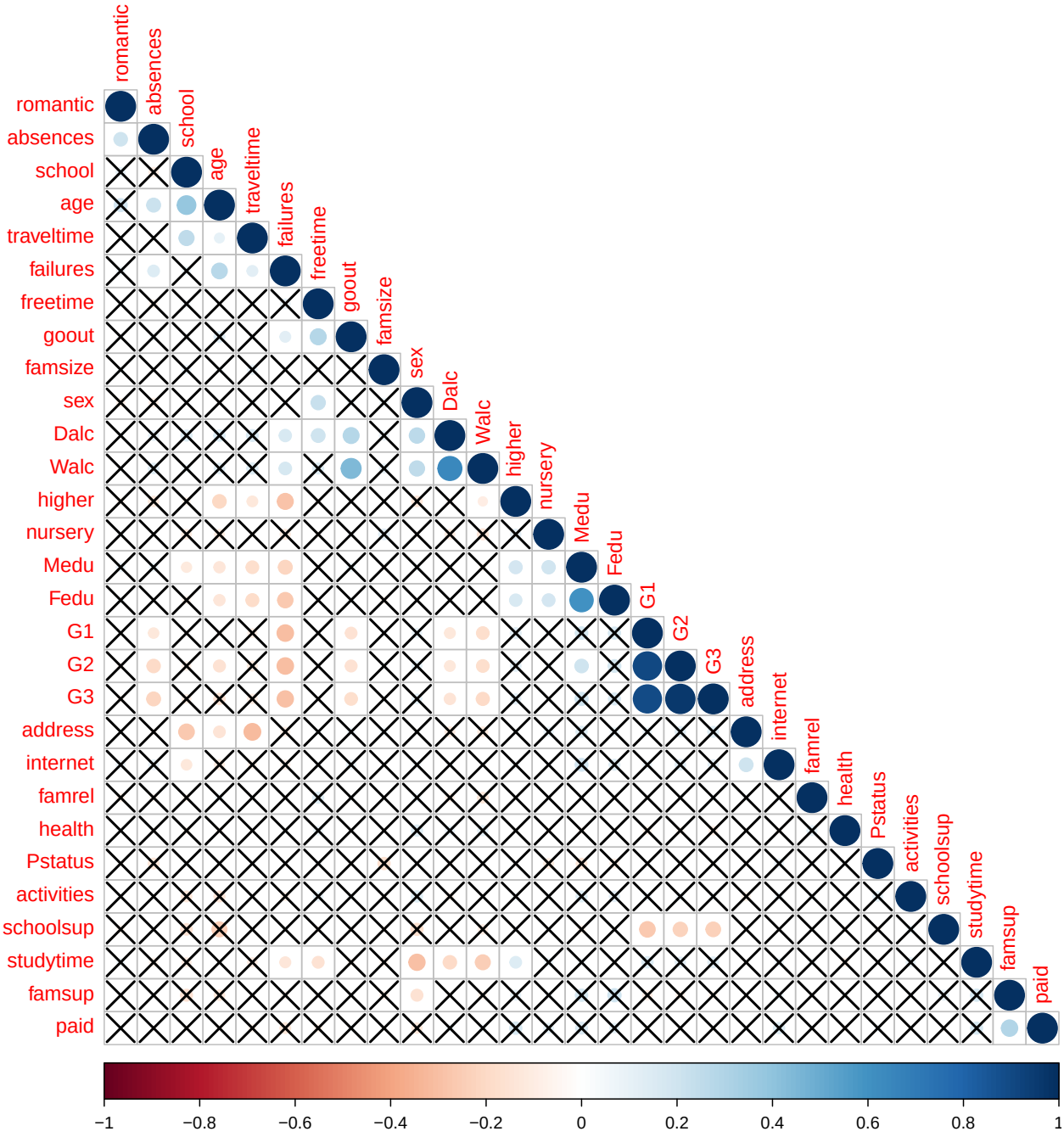


Plot 2

```
### binary column names
bin_col <- c('school','sex','address','famsize','Pstatus','schoolsup','famsup',
             'paid','activities','nursery','higher','internet','romantic')
num_col <- c('age','Medu','Fedu','traveltime','studytime','failures','famrel',
             'freetime','goout','Dalc',"Walc",'health','absences',"G1","G2","G3")
## turn the data frame into all numeric variables.
df_fact <-
  as.data.frame(unclass(df %>% select(c(bin_col, num_col))), stringsAsFactors = T)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(bin_col)` instead of `bin_col` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(num_col)` instead of `num_col` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
df_fact <- lapply(df_fact, as.numeric) %>% as_tibble()
## calculate correlation and significance, and draw the correlation plot.
sigtest <- cor.mtest(cor(df_fact), conf.level = 0.95)
new_p2 <- corrplot(
  cor(df_fact),
  order = 'hclust',
  p.mat = sigtest$p,
  sig.level = 0.05,
  type = 'lower'
)
```



Discussion

1. Why did you choose the data that you did? What makes it interesting to you?

The data I choose contains the scores at the end of a math program with several features of each students. Among the features, several that relate to the students' familial background are notably interesting, including education attainment of mother and father. I've been investigating the complex interplay between one's family environment and innate skill's effects on one's trait. Therefore, while understanding the lack of representativeness of the dataset,

2. What are some interesting facts that you learned through the visualizations?

Educational attainment of parents are classified as:

0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education

Fig 1. One interesting finding through the visualization would be that the highest parental educational attainment is mildly associated with the student's grade, but the differences in the "variation of the final grade" between each group is large. As the parental education increase, the variation increase.

Fig 2. Only the significant correlations at $\alpha = 0.05$ are displayed with circles. The insignificant ones are all filled by a X. From the correlation plot, we can see that the mother's education is significantly and positively correlated with one of the grade (G2), but father's education is not significantly correlated with any grades.

Also, we can see a significant positive correlation between mother's education and father's education, suggesting the possibility of assortative mating, which means the people with similar phenotypes are more likely to mate with each other.

3. How did you create the plot? Were there any data preparation steps you used? What guided the style customization you used?

For both plot, I removed 38 students with 0 final grade.

For Fig 1, I used `mutate()` to get the variable that represent the highest educational attainment among the parents of the student. I first plot boxplots of the grade in different parental education group. Then, I plot the grade of the students on y-axis and the highest parental education on x-axis. The points are jittered in both x and y directions to avoid overlapping. They are colored by the school that student goes to.

For the Fig 2, I only used binary variables and numeric variables. I turned them all into numeric variables to calculate the correlation matrix. Then use corrplot() function and cor.mtest() function to estimate correlations and test for significance.

I customized the color palettes for the plots I draw. One useful package is "ggsci", which assign colors according to the color used in scientific journals. In this portfolio, I used the colors from American Association for the Advancement of Science.