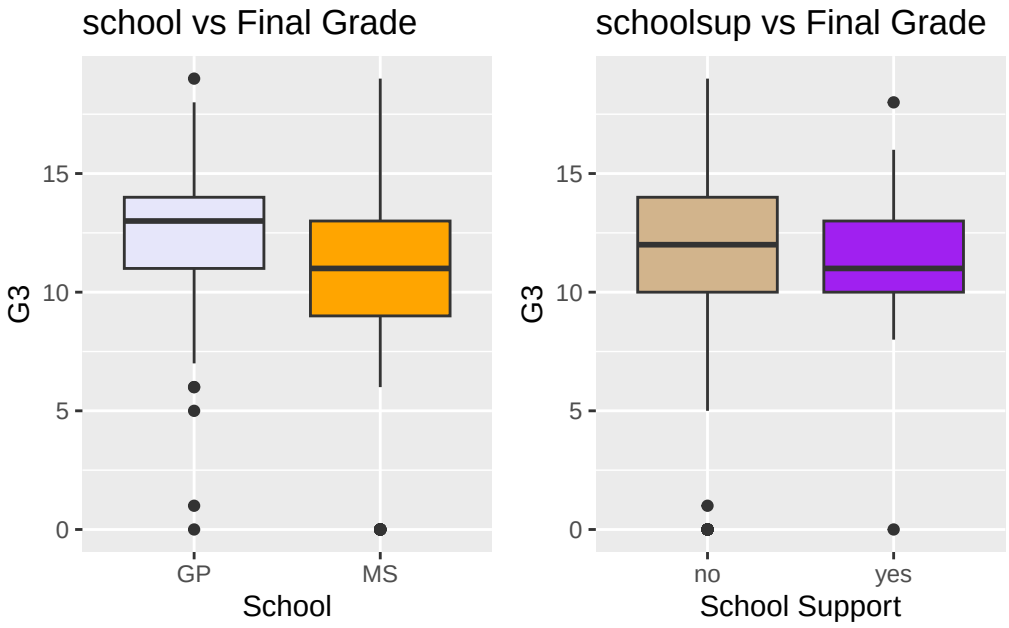
PSTAT 100: Step 2

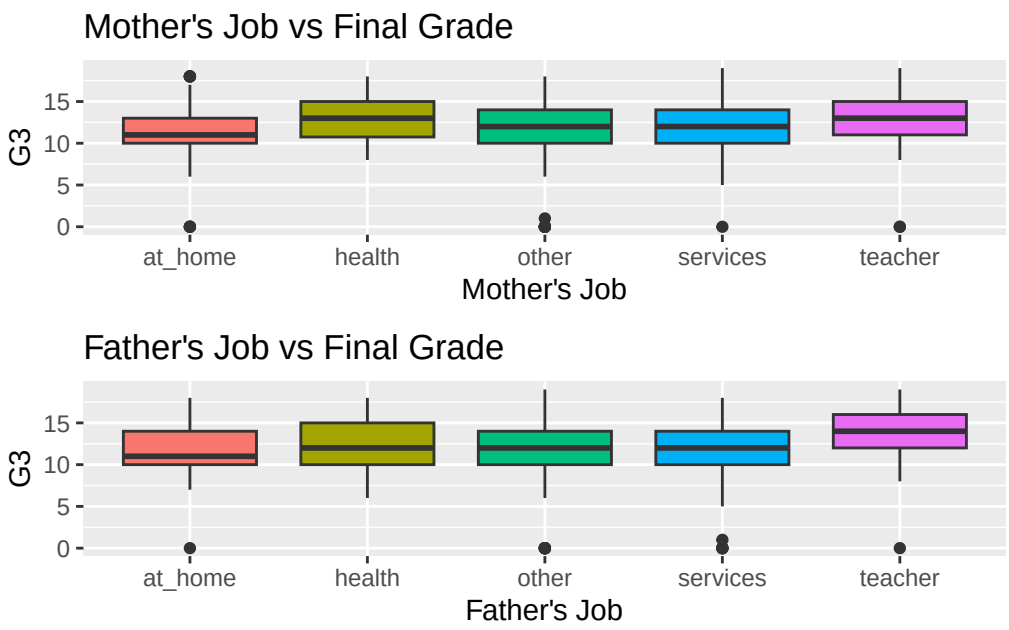
Darren Colianni

1 Visualize: (Study Time, Parent Edu, Internet) -> Final Grade

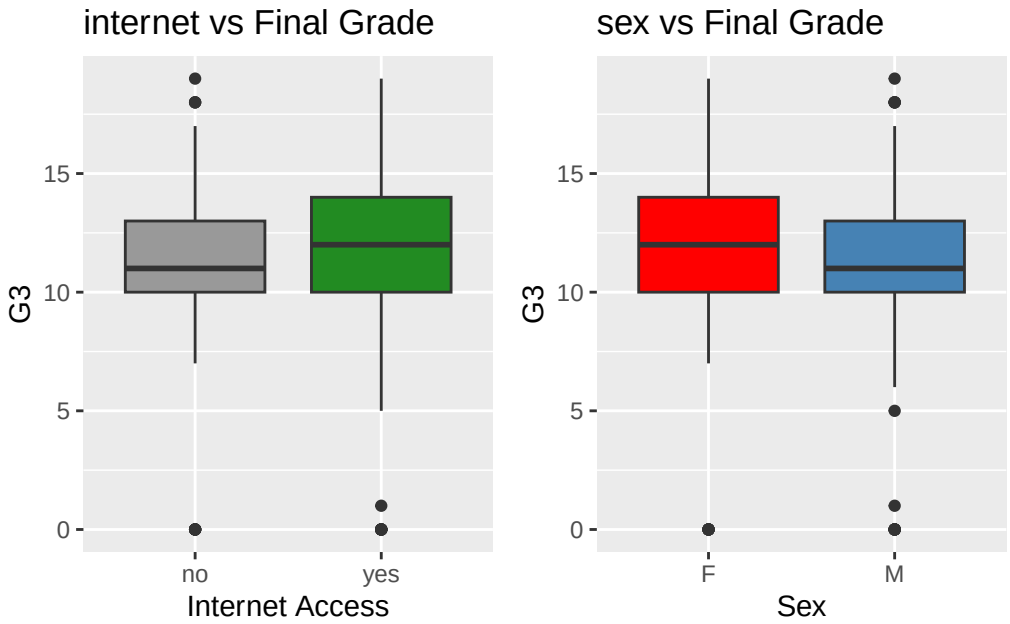
```
p1 <- make_boxplot(df, "school", c("GP" = "lavender", "MS" = "orange"), xlabel = "School")
p2 <- make_boxplot(df, "schoolsup", c("yes" = "purple", "no" = "tan"), xlabel = "School Suppo
p3 <- make_boxplot(df, "Mjob", title = "Mother's Job vs Final Grade", xlabel = "Mother's Job
p4 <- make_boxplot(df, "Fjob", title = "Father's Job vs Final Grade", xlabel = "Father's Job
p5 <- make_boxplot(df, "internet", c("yes" = "forestgreen", "no" = "gray60"), xlabel = "Inter
p6 <- make_boxplot(df, "sex", c("F" = "red", "M" = "steelblue"), xlabel = "Sex")

p1 | p2
```



```
p3 / p4
```


Mother's Job vs Final Grade
G3
15
10
5
0
at_home
health
other
services
teacher
Mother's Job
Father's Job vs Final Grade
G3
15
10
5
0
at_home
health
other
services
teacher
Father's Job

```
p5 | p6
```

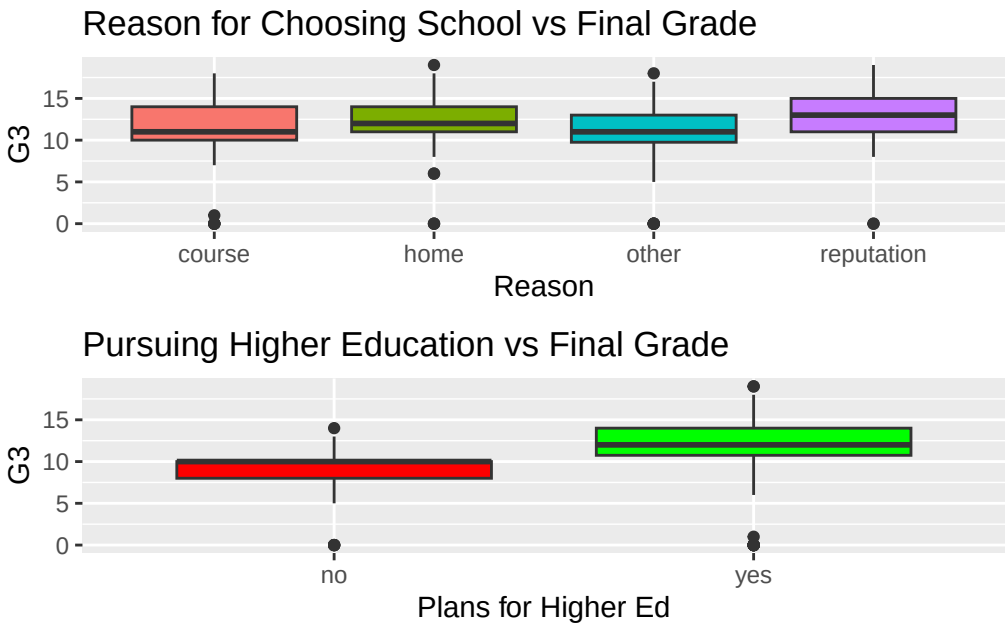

internet vs Final Grade
G3
15
10
5
0
no
yes
Internet Access
sex vs Final Grade
G3
15
10
5
0
F
M
Sex

The GP school has a higher median G3 among the two. School support indicates struggling students–its median is slightly lower than those not receiving support.

For parent professions, the 95% median line indicates teacher parent profession is highest, and at home is lowest. Health is highest with teacher for mother's, and in the middle range for father's.

Having internet access is significant as well, and so is the student's sex: having internet and being female having 95% confidence medians above the opposite groups.

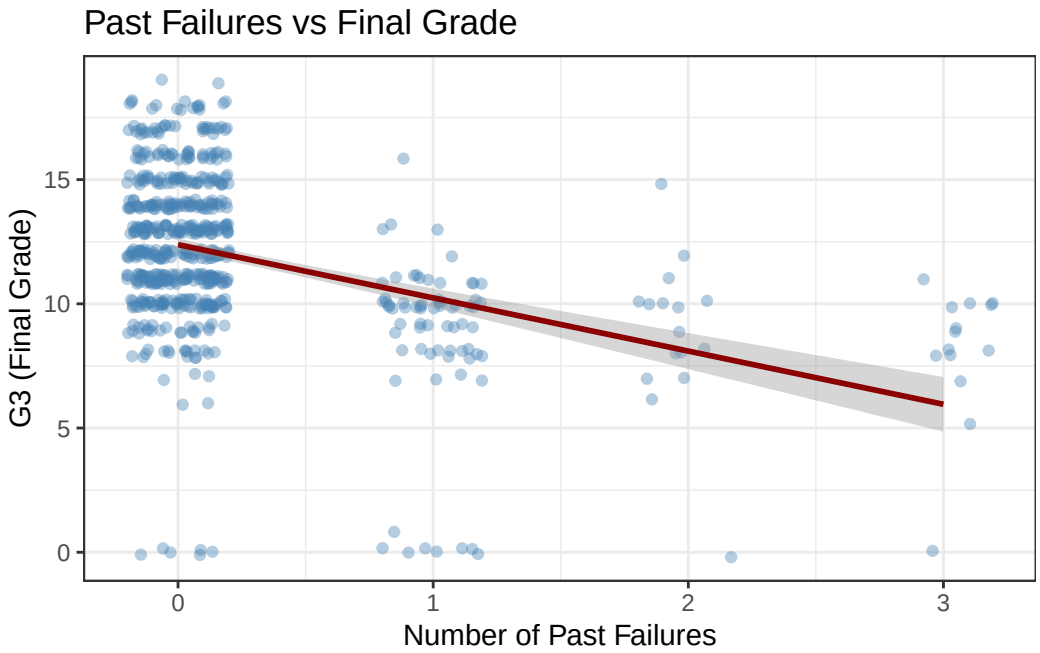
```
p_reason <- make_boxplot(df, "reason",
                         title = "Reason for Choosing School vs Final Grade",
                         xlabel = "Reason")

p_higher <- make_boxplot(df, "higher", fill_colors = c("yes" = "green", "no" = "red"),
  title = "Pursuing Higher Education vs Final Grade",
  xlabel = "Plans for Higher Ed")

p_reason / p_higher
```



Both of these factors indicate motivation. So, as is evident, they have significantly different medians between each of these two sets of groups.

```r
ggplot(df, aes(failures, G3)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "lm", color = "darkred") +  # se = TRUE is default
  labs(title = "Past Failures vs Final Grade", x = "Number of Past Failures", y = "G3 (Final
  theme_bw()
```

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



This is the strongest predictor that is not based on G2 grade. Its R^2 is rather weak, but bt tracing a horizontal line from the left to the right, finding if the left's shaded bottom is above a point on the right's, one finds: - 0 failures has a significantly greater predicted final grade than the rest - 1 failure has a significantly greater predicted final grade than 3 failures.