Final project

Sophie Perez

Load in the data and view it

```
scores <- read.csv("student-por.csv")
knitr::kable(head(scores, 10))
```

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	no	no	yes	yes	yes	no	4	2	2	1	1	1	0	15	16	17
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	no	yes	yes	yes	yes	no	5	5	1	1	1	5	0	12	12	13

Libraries we will use

```
library(ggplot2)
library(ggcorrplot)
```

```
original_len <- nrow(scores)
scores <- na.omit(scores)
final_len <- nrow(scores)
cat("Removed", original_len-final_len, "rows with NA values")
```

```
## Removed 0 rows with NA values
```

Remove the features I don't want to use *mention why they were not used

```
unused_features <- c("school", "sex", "address", "famsize", "Medu", "Fedu", "Mjob", "Fjob", "guardian",
for (feature in unused_features) {
  scores[feature] <- NULL
}
knitr::kable(head(scores, 10))
```

age	Pstatus	reason	traveltime	studytime	failures	schoolsup	famsup	paid	activities	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
18	A	course	2	2	0	yes	no	no	no	yes	no	no	4	3	4	1	1	3	4	11	11	11
17	T	course	1	2	0	no	yes	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
15	T	other	1	2	0	yes	no	no	no	yes	yes	no	4	3	2	2	3	3	6	12	13	12
15	T	home	1	3	0	no	yes	no	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
16	T	home	1	2	0	no	yes	no	no	yes	no	no	4	3	2	1	2	5	0	11	13	13
16	T	reputation	1	2	0	no	yes	no	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13

age	Pstatus	reason	traveltime	studytime	failures	schoolsup	famsup	paid	activities	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G3
16	T	home	1	2	0	no	no	no	no	yes	yes	no	4	4	4	1	1	3	0	13
17	A	home	2	2	0	yes	yes	no	no	yes	no	no	4	1	4	1	1	1	2	13
15	A	home	1	2	0	no	yes	no	no	yes	yes	no	4	2	2	1	1	1	0	17
15	T	home	1	2	0	no	yes	no	yes	yes	yes	no	5	5	1	1	1	5	0	13

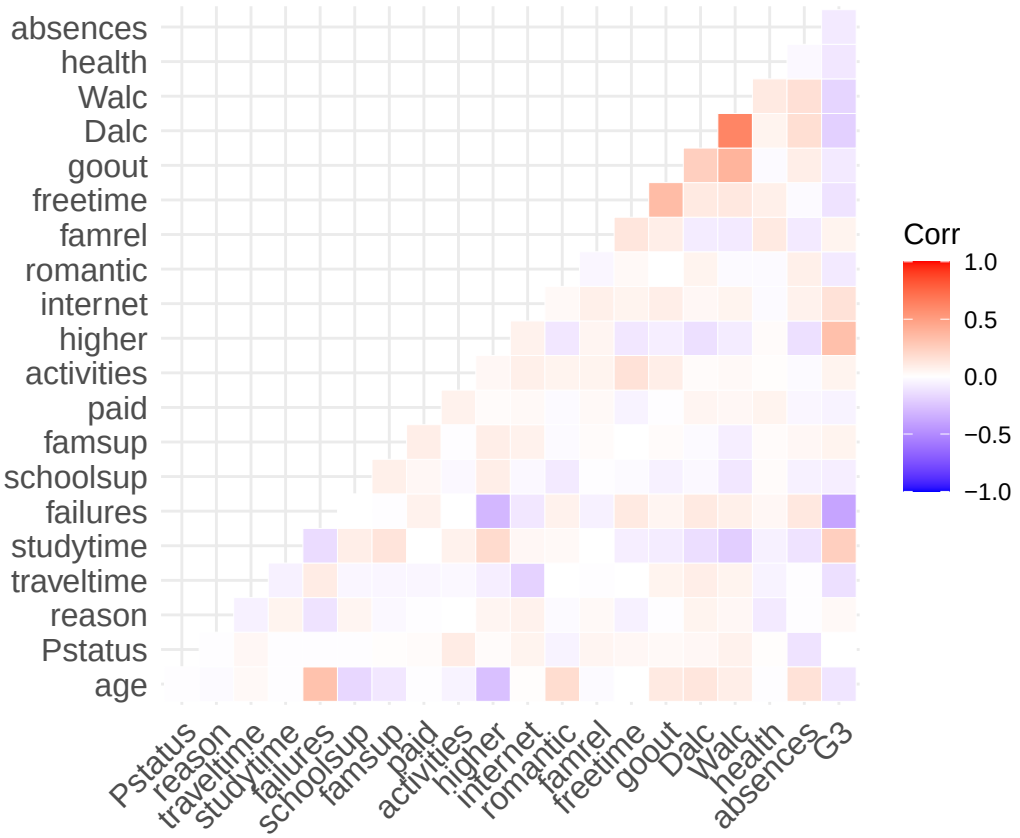
Convert non-numeric features to numeric

```
scores$Pstatus <- ifelse(scores$Pstatus == "A", 0, 1)

scores$reason <- ifelse(scores$reason == "course", 0,
                        ifelse(scores$reason == "home", 1,
                               ifelse(scores$reason == "reputation", 2, 3)))

binaries <- c("schoolsup", "famsup", "paid", "activities", "higher", "internet", "romantic")
for (feature in binaries) {
  scores[feature] <- ifelse(scores[feature] == "yes", 1, 0)[,1]
}

knitr::kable(head(scores, 10))
```

age	Pstatus	reason	traveltime	studytime	failures	schoolsup	famsup	paid	activities	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G3
18	0	0	2	2	0	1	0	0	0	1	0	0	4	3	4	1	1	3	4	11
17	1	0	1	2	0	0	1	0	0	1	1	0	5	3	3	1	1	3	2	11
15	1	3	1	2	0	1	0	0	0	1	1	0	4	3	2	2	3	3	6	12
15	1	1	1	3	0	0	1	0	1	1	1	1	3	2	2	1	1	5	0	14
16	1	1	1	2	0	0	1	0	0	1	0	0	4	3	2	1	2	5	0	13
16	1	2	1	2	0	0	1	0	1	1	1	0	5	4	2	1	2	5	6	13
16	1	1	1	2	0	0	0	0	0	1	1	0	4	4	4	1	1	3	0	13
17	0	1	2	2	0	1	1	0	0	1	0	0	4	1	4	1	1	1	2	13
15	0	1	1	2	0	0	1	0	0	1	1	0	4	2	2	1	1	1	0	17
15	1	1	1	2	0	0	1	0	1	1	1	0	5	5	1	1	1	5	0	13

Now that the data is all numeric, we can look at the correlations

```
correlations_matrix <- cor(scores)

ggcorrplot(correlations_matrix, type = "lower", outline.color = "white", colors = c("blue", "white", "r
```



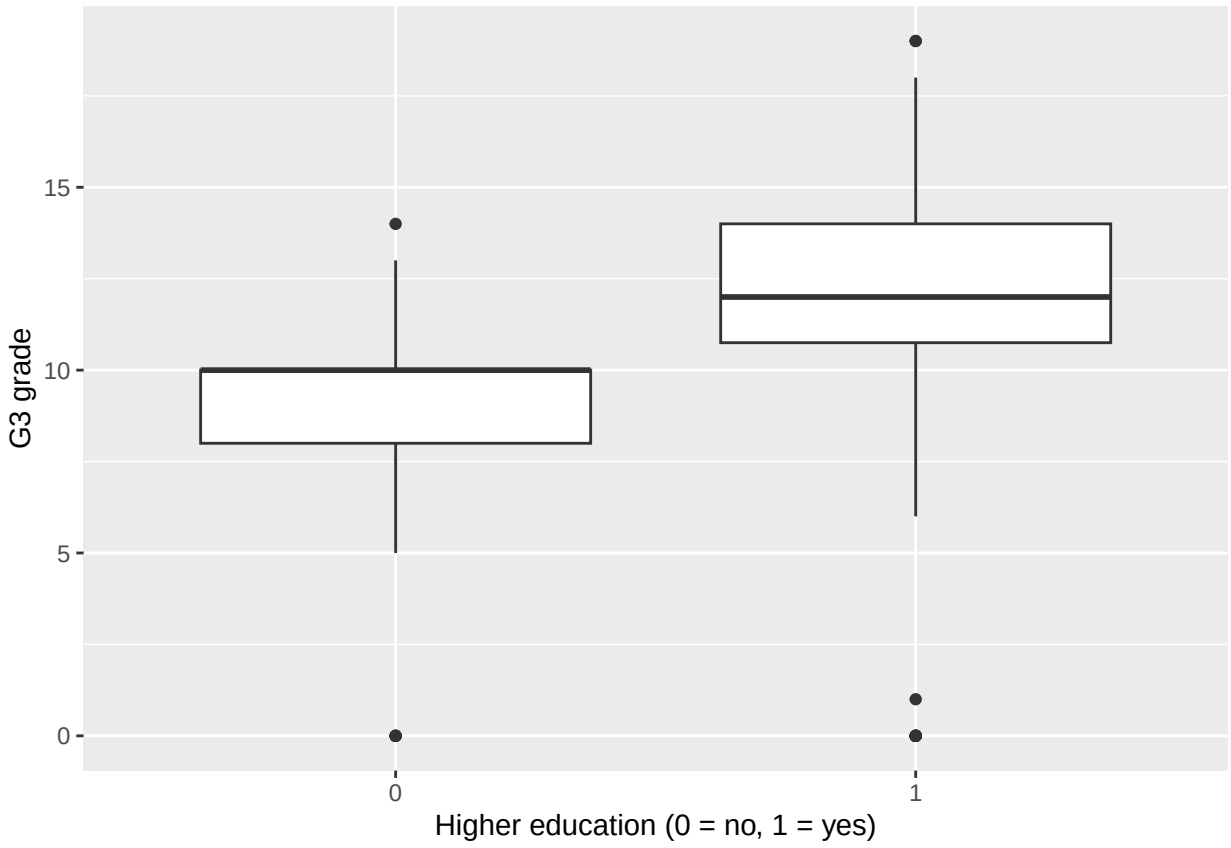
This heat graph shows us that surprisingly, there is not much of a correlation on alcohol consumption. and a students performance; answering my question. This visualization does give more information on what is tied to better academic performance though (G3 aka Final Scores), i.e., wanting to pursue higher education, age, more study time

There are some higher correlations that we can ignore: Walc vs Dalc (weekend and weekday drinking habits don't differ much for a person), Walc vs goout (an obvious correlation: if you go out more on weekends, you'll likely drink more)

The correlations that stand out are: higher vs failures (negative), G3 vs failures (negative), failures vs age (positive),
higher vs G3 (positive)

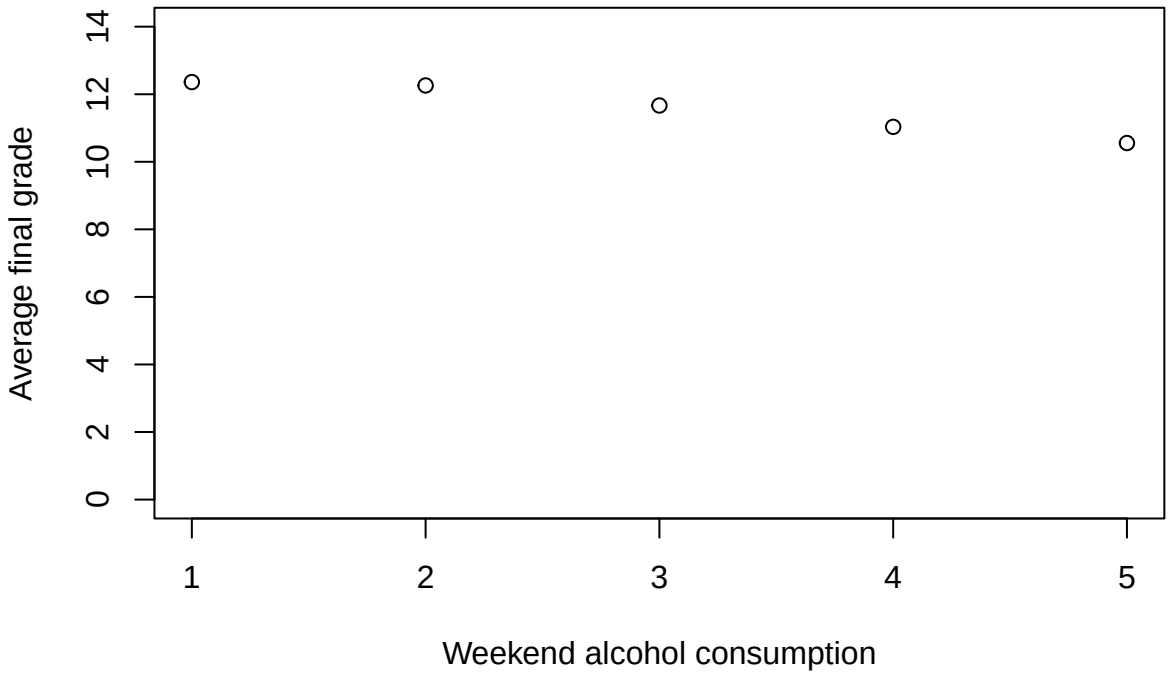
The feature that had the strongest correlation with the final grade is the **higher** feature, which is a boolean value that represent if the person plans to pursue a higher education after college.

```
ggplot(scores, aes(x = factor(higher), y = G3)) +
  geom_boxplot() +
  labs(x = "Higher education (0 = no, 1 = yes)", y = "G3 grade")
```



Box plots showing students students who wish to pursure higher education and those who dont; along with their average final scores (G3). This showcasees that those who wish to attend higher education also tend to perform better in school.

```
averages <- aggregate(scores$G3, by=list(Walc=scores$Walc), FUN=mean)

plot(averages$Walc, averages$x, ylim = c(0,14), ylab="Average final grade", xlab="Weekend alcohol consu
```



Average final grades (G3 score) per weekly alcohol consumption

```
averages <- aggregate(scores$G3, by=list(higher=scores$higher), FUN=mean)

cat("Average score for those not pursuing higher education:", averages[1,2], "\n")
```

```
## Average score for those not pursuing higher education: 8.797101
```

```
cat("Average score for those pursuing higher education:", averages[2,2])
```

```
## Average score for those pursuing higher education: 12.27586
```