Lab 2

Christopher Hainzl, Christopher Hakkenberg, Michael Ayaz

2023-02-24

```
library('tidyverse')
```

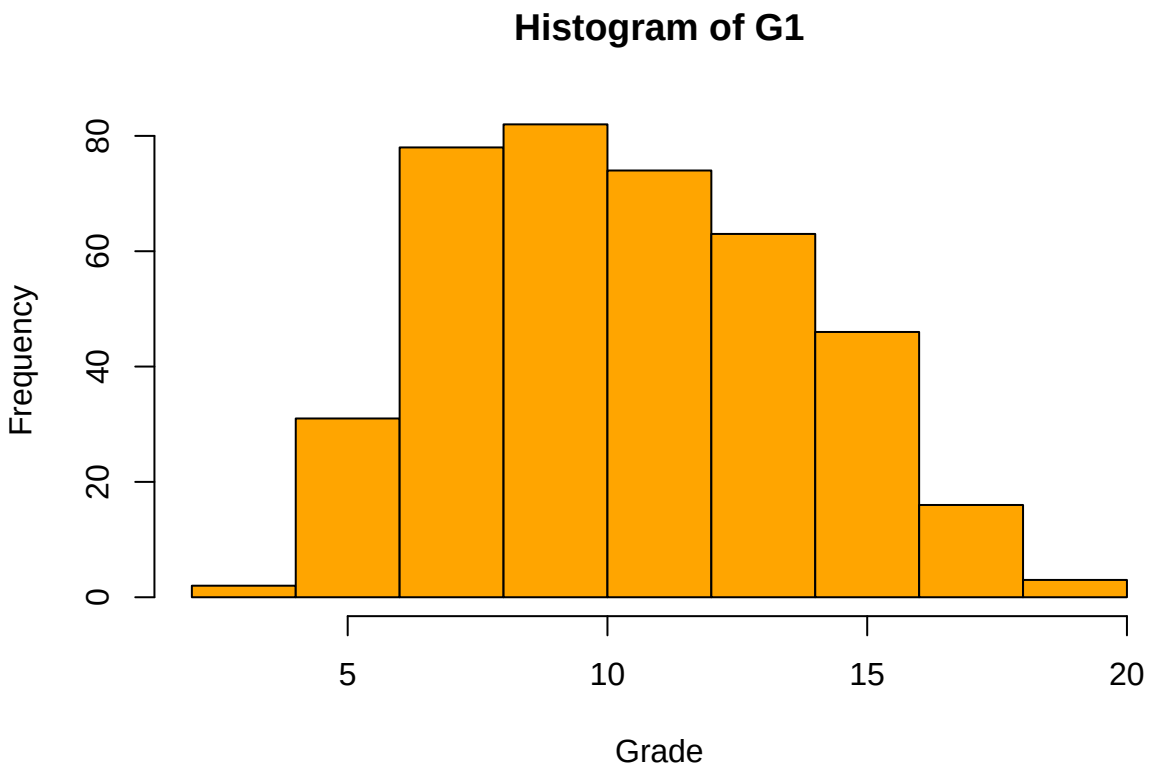
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

We will be using the dataset named "studentdata" to try to determine which variables will be most useful in predicting a student's final grade: alcohol consumption, or extra paid classes.

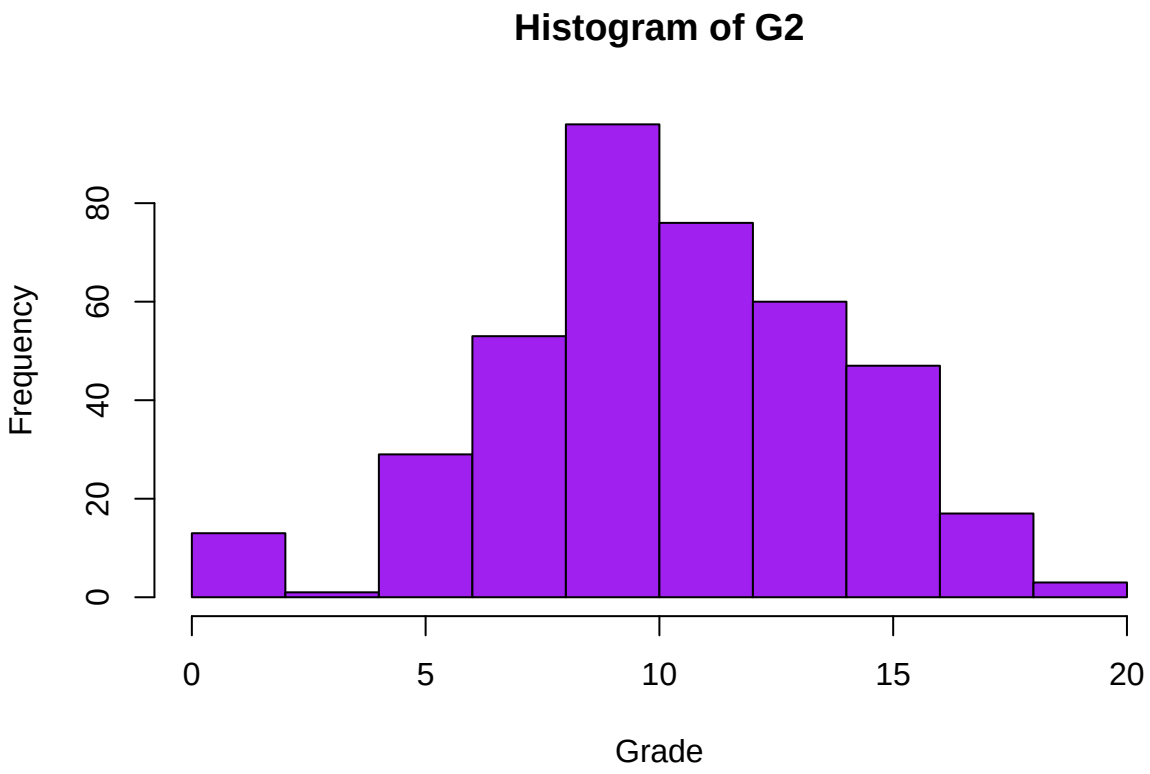
To prepare our data, we will read in the dataset using the read.csv() function and filter out the students who finished with a final grade of 0. The reason we are choosing to filter out those students is because they might have dropped out before the school year finished up. We are only concerned with students who made it to the end of the school year. We also do not want the values equal to 0 to affect our statistical measurements.

```
studentdata = read.csv("student-mat(1).csv", sep = ";")
```

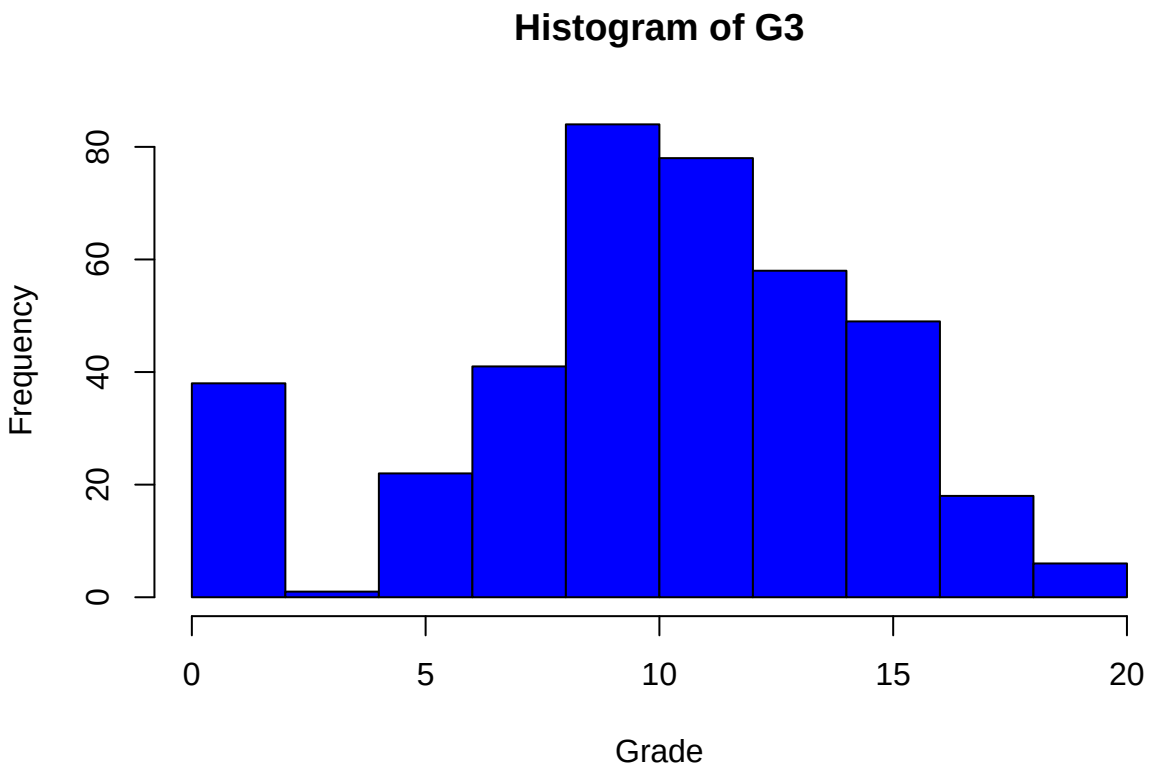
```
hist(studentdata$G1, main = "Histogram of G1", xlab = "Grade", col = "orange")
```




```
hist(studentdata$G2, main = "Histogram of G2", xlab = "Grade", col = "purple")
```

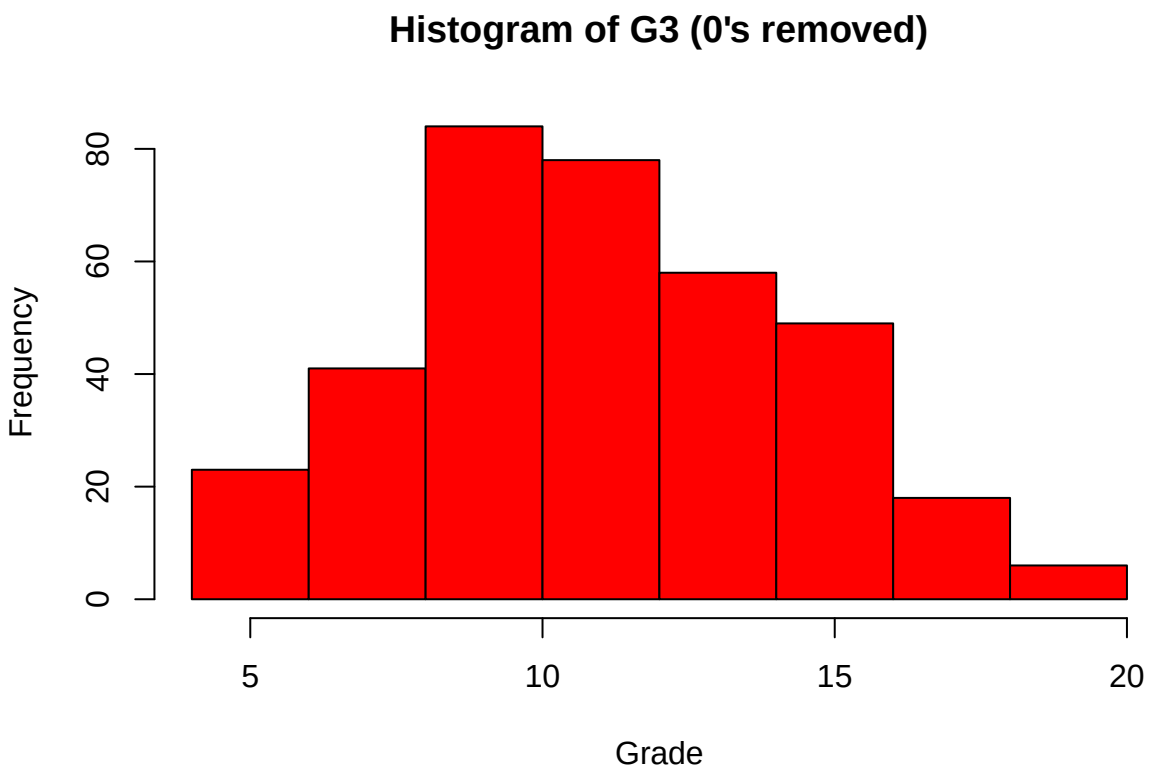



```
hist(studentdata$G3, main = "Histogram of G3", xlab = "Grade", col = "blue")
```




```
studentdata <- studentdata %>% filter(G3 != 0)
```

```
hist(studentdata$G3, main = "Histogram of G3 (0's removed)", xlab = "Grade", col = "red")
```



After removing the final grades equal to 0, the histogram of the final grades appears to follow the bell curve. This means the data is normally distributed.

```
t.test(studentdata$G3, alternative = "greater", mu = 10, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  studentdata$G3
## t = 8.9199, df = 356, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 10
## 95 percent confidence interval:
##  11.24208      Inf
## sample estimates:
## mean of x
##  11.52381
```

Since our p-value is less than .05, we reject the null hypothesis and say there is statistically significant evidence that the average final grade is significantly greater than 10.

```
t.test(studentdata$G3, alternative = "two.sided", conf.level = .95)
```

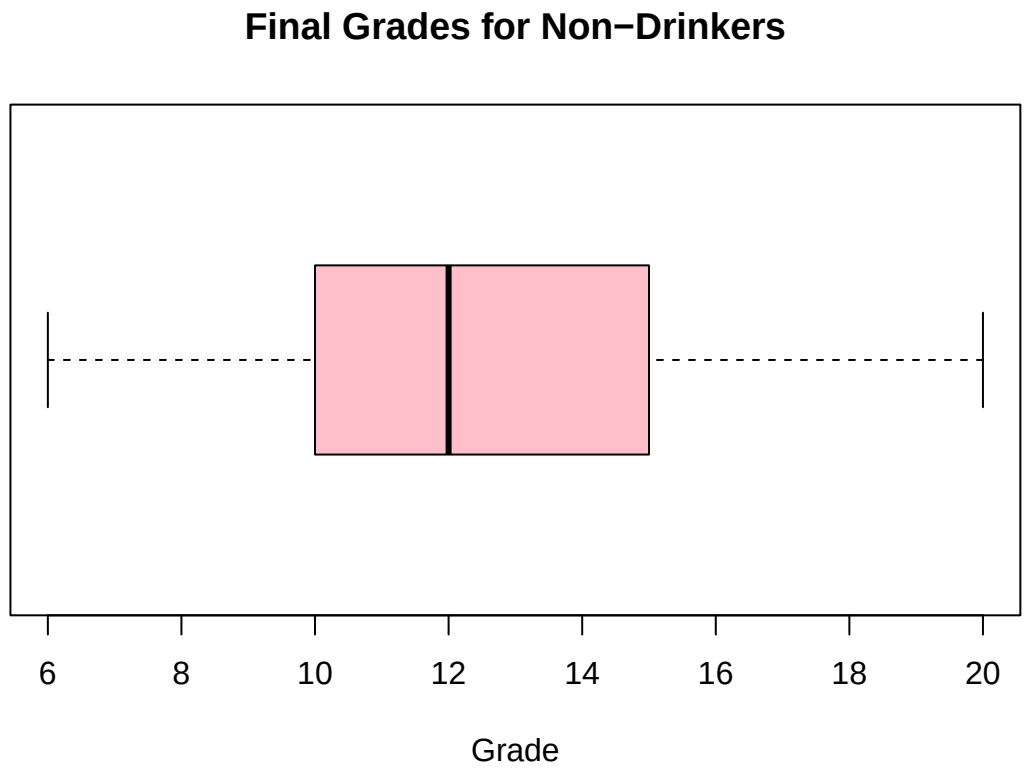
```
##
##  One Sample t-test
##
## data:  studentdata$G3
## t = 67.457, df = 356, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  11.18784 11.85978
## sample estimates:
## mean of x
##  11.52381
```

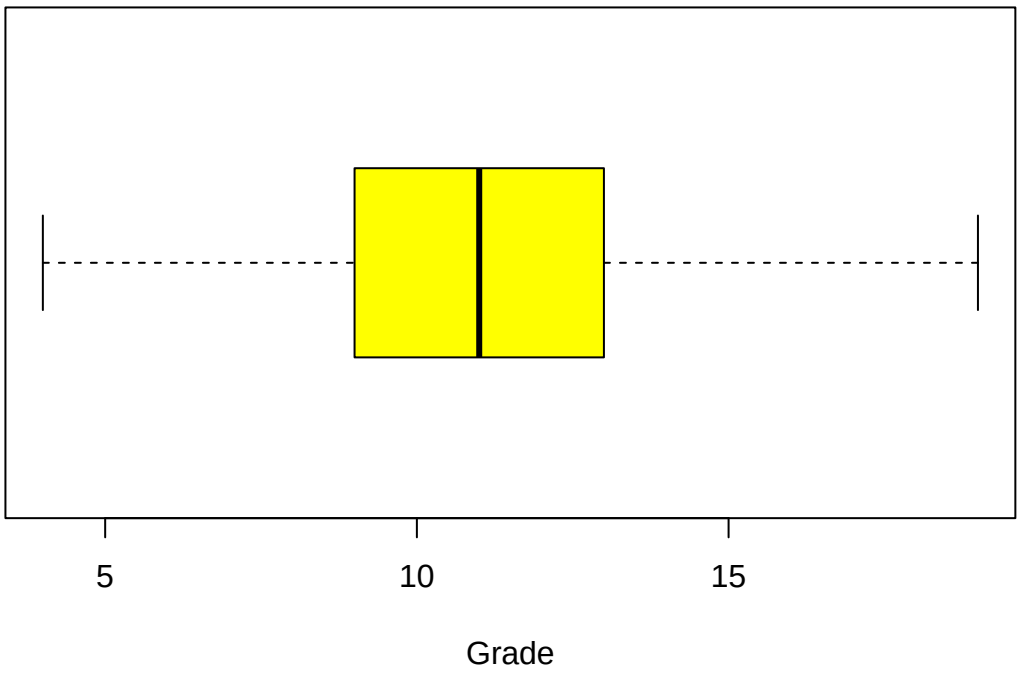
The 95% confidence interval for the true mean final grade is approximately 11.2 < mu < 11.9. The sample mean falls within this range.

```
noalcohol <- studentdata %>% filter(Walc == 1)
alcohol <- studentdata %>% filter(Walc > 1)
```

```
boxplot(noalcohol$G3, horizontal = TRUE, xlab = "Grade", col = "pink", main = "Final Grades for Non-Dri
```




```
boxplot(alcohol$G3, horizontal = TRUE, xlab = "Grade", col = "yellow", main = "Final Grades for Drinker
```




```
t.test(noalcohol$G3, alcohol$G3, alternative = "two.sided", conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  noalcohol$G3 and alcohol$G3
## t = 2.9198, df = 246.38, p-value = 0.003827
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3444637 1.7725475
## sample estimates:
## mean of x mean of y
##  12.18797  11.12946
```

Since our p-value is less than .05, we reject the null hypothesis and say there is significant evidence that there is a difference in mean final grades between drinkers and non-drinkers. The sample mean for the non-drinkers was approximately 12.19, and the sample mean for the drinkers was approximately 11.13. This makes sense considering that alcohol impairs the functionality of the brain.

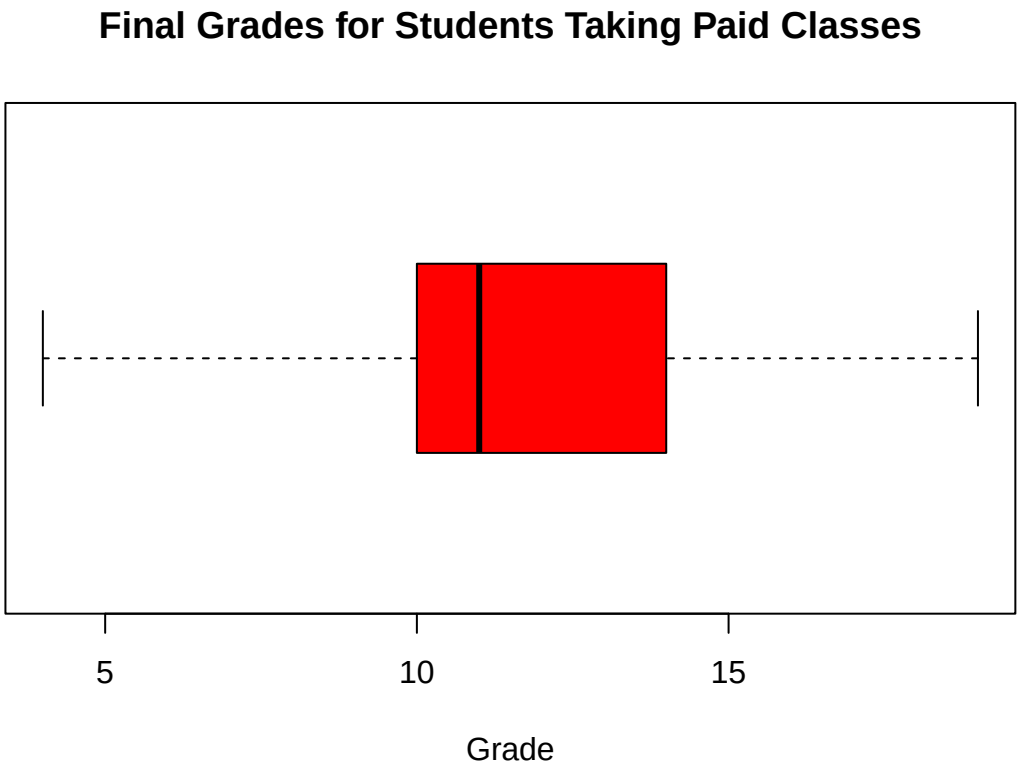
```
paid <- studentdata %>% filter(paid == "yes")
notpaid <- studentdata %>% filter(paid == "no")
```

```
length(paid) / (length(notpaid) + length(paid))
```

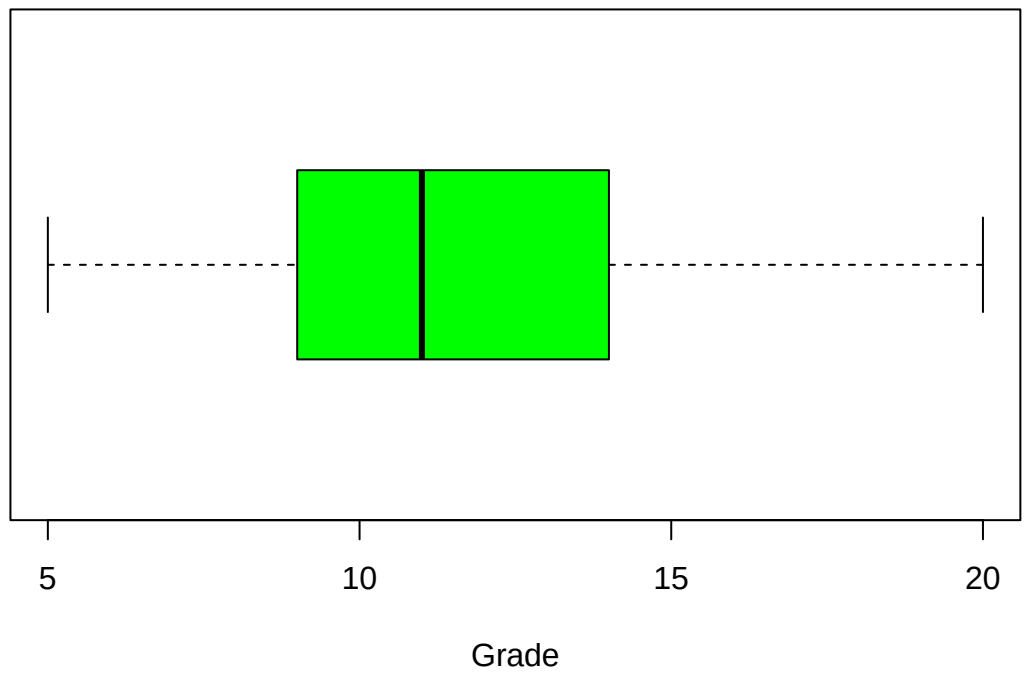
```
## [1] 0.5
```

It turns out that approximately half of all the students utilize extra paid classes.

```
boxplot(paid$G3, horizontal = TRUE, xlab = "Grade", col = "red", main = "Final Grades for Students Taki
```




```
boxplot(notpaid$G3, horizontal = TRUE, xlab = "Grade", col = "green", main = "Final Grades for Students
```




```
t.test(notpaid$G3, paid$G3, alternative = "two.sided", conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  notpaid$G3 and paid$G3
## t = 0.54661, df = 354.09, p-value = 0.585
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4842183  0.8569878
## sample estimates:
## mean of x mean of y
##  11.61413  11.42775
```

Since the p-value is greater than .05, we do not reject the null hypothesis. There is not significant evidence of a difference in average final grades between students who took paid classes and those who did not. In addition to the p-value, this is also supported by the fact that the average final grade for the students who took paid classes was approximately 11.61, and the average final grade for those who did not was approximately 11.43. Because of this, we find it surprising that about half of all the students are taking extra paid classes. However, the lower average final grade for students taking extra paid classes may have been due to some of those students having a Walc value greater than 1 (thus impairing their academic performance).

In conclusion, based on our boxplots and hypothesis tests, we believe that weekend alcohol consumption (Walc) could be more useful in predicting a student's final grade than whether or not the student is taking extra paid classes (since the hypothesis test for alcohol consumption resulted in a much lower p-value than the one for extra paid classes).