Mock Coding Challenge: Self-Assessment Guide

ADS2

Semester 1, 2023/24

Instructions

Using this guide, you should be able to go through your mock coding challenge and assess how you did. Use this to figure out how many points you would have gotten for each part, where you did well, and where you could have done better. Make sure to make notes that will help you do well on the actual assessment in January.

1. Stroop test [25 points in total]

Import the data and plot them in a useful way. [6 points]

Check for the following:

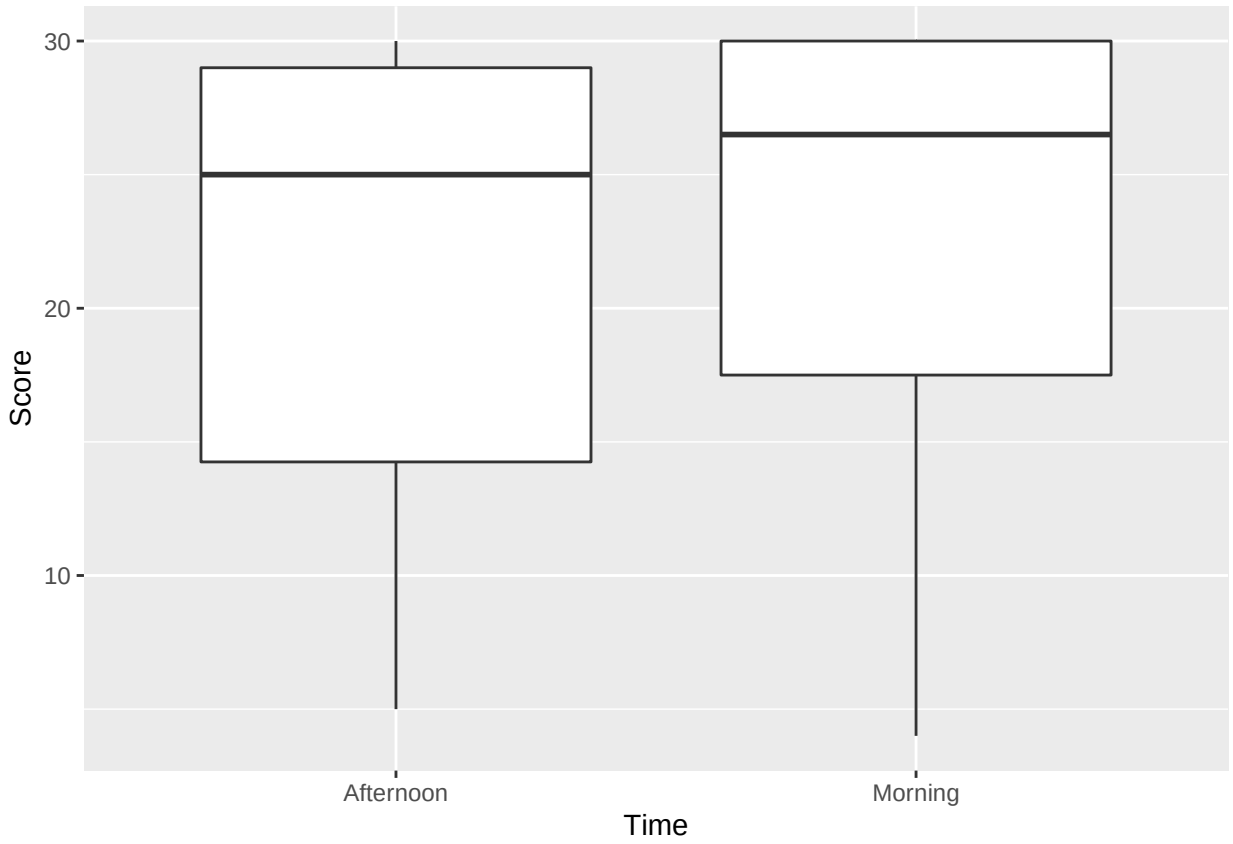
☐ Has the data been successfully imported? We do not require you to show your code here (since it may contain identifying information if your file path contains your name). But you should have done something like this:

```
stroop <- read.csv("stroop_test.csv", header=TRUE)
head(stroop)
```

```
##      Time Score
## 1 Morning    14
## 2 Morning    21
## 3 Morning    30
## 4 Morning    30
## 5 Morning    26
## 6 Morning    19
```

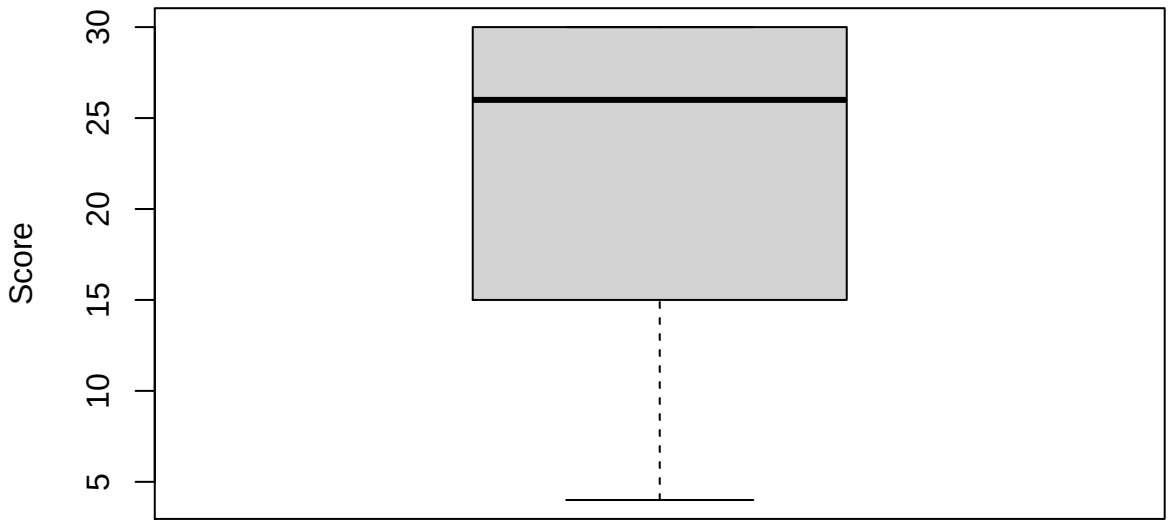
☐ Has the data been plotted in a useful way? Here is a useful plot:

```
library(ggplot2)
p <- ggplot(stroop,aes(x=Time,y=Score)) + geom_boxplot()
p
```



Why is this kind of plot useful? First, because it shows both groups. The plot below would get some points (because a plot has been produced), but would be less useful, because it does not distinguish between the "afternoon" and "morning" group:

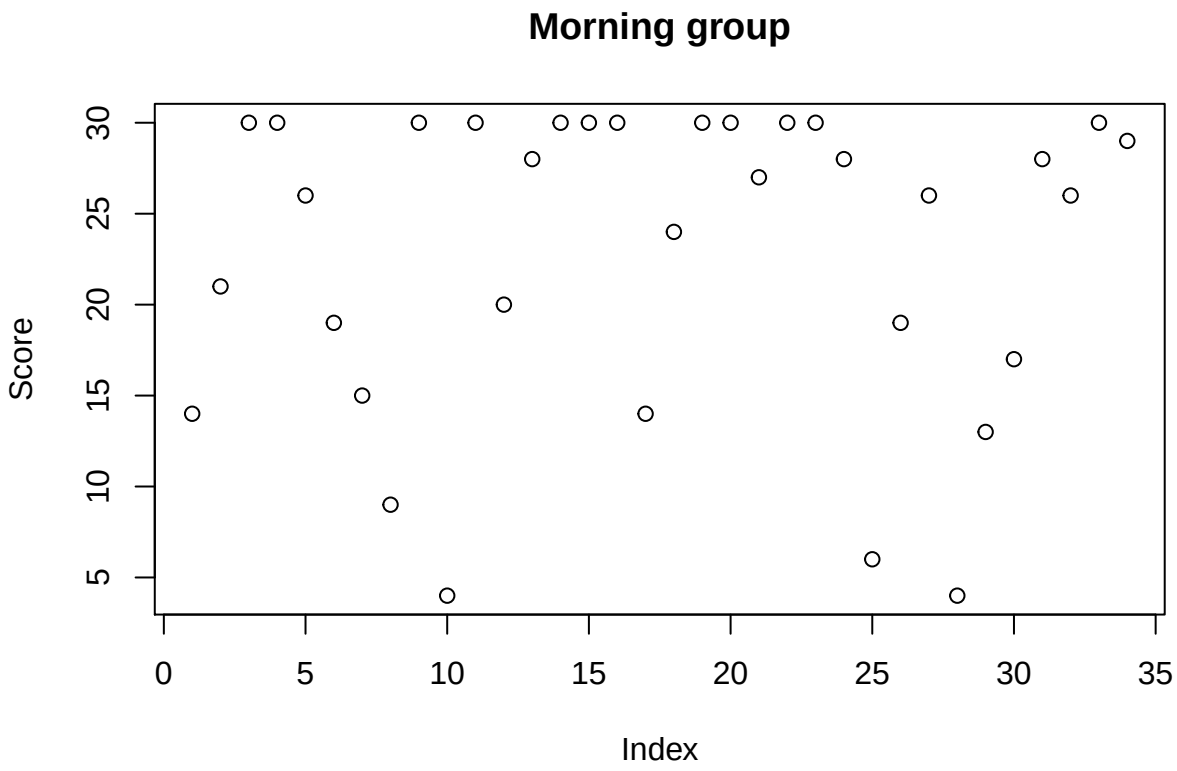
```
boxplot(stroop$Score,ylab="Score")
```



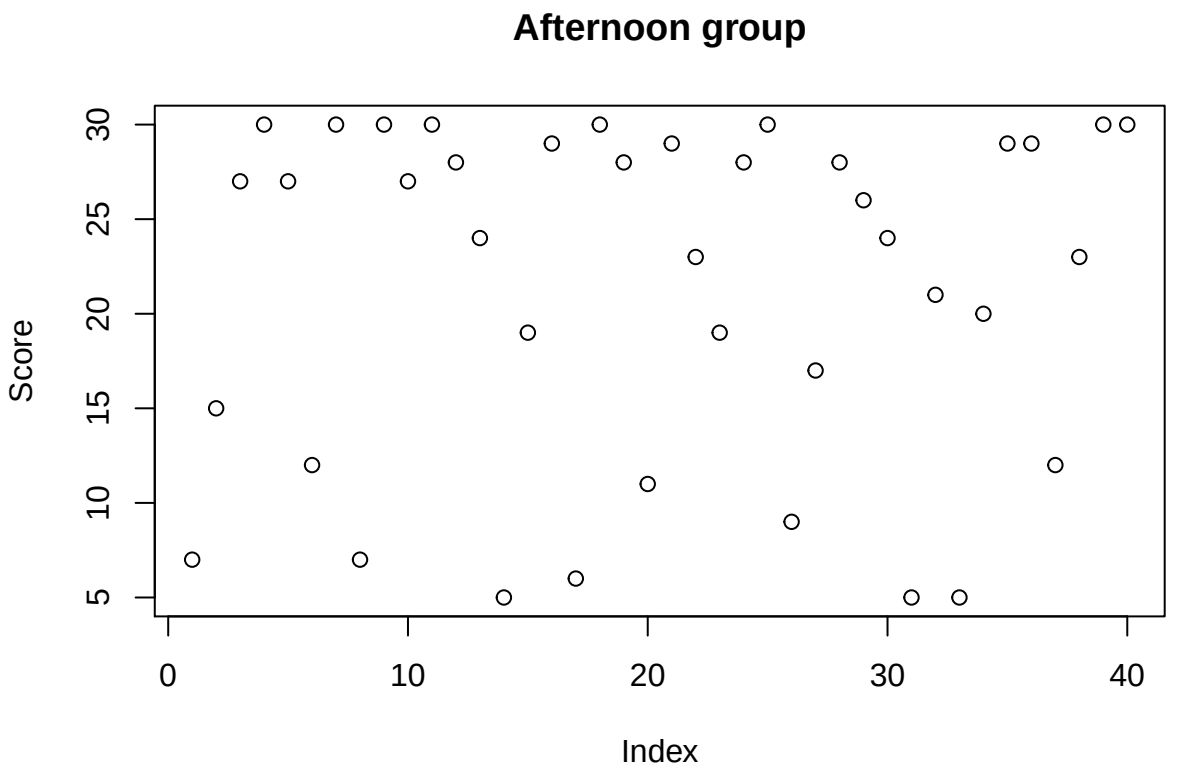
The other reason why the first plot is useful is that it gives an idea of the underlying distribution. For instance, we can see the medians, ranges, and IQRs of both groups. Other plot formats are also fine, as long as they similarly show the distribution. For instance, you could show the histograms of both groups. (Also you don't have to do them both in one plot; it's fine to have one plot for the Morning group and one for the Afternoon group.)

BUT the following kind of plot, for instance, is not very useful:

```
plot(stroop[stroop$Time=="Morning", "Score"], ylab="Score", main="Morning group")
```




```
plot(stroop[stroop$Time=="Afternoon", "Score"], ylab="Score", main="Afternoon group")
```



These plots would get partial points (because something has been plotted), but not full points. The reason is that the x axis does not carry any useful information - it is just the order in which the samples appear in the data set. The space would be better used for other types of plots, as discussed above.

☐ Are all axes clearly labelled? If a plot uses two colours, is it clear to the reader what the colours mean?

Is there a difference in performance on the Stroop task between the morning and afternoon group? [14 points]

This is clearly the main and most important part of this question. Check for the following

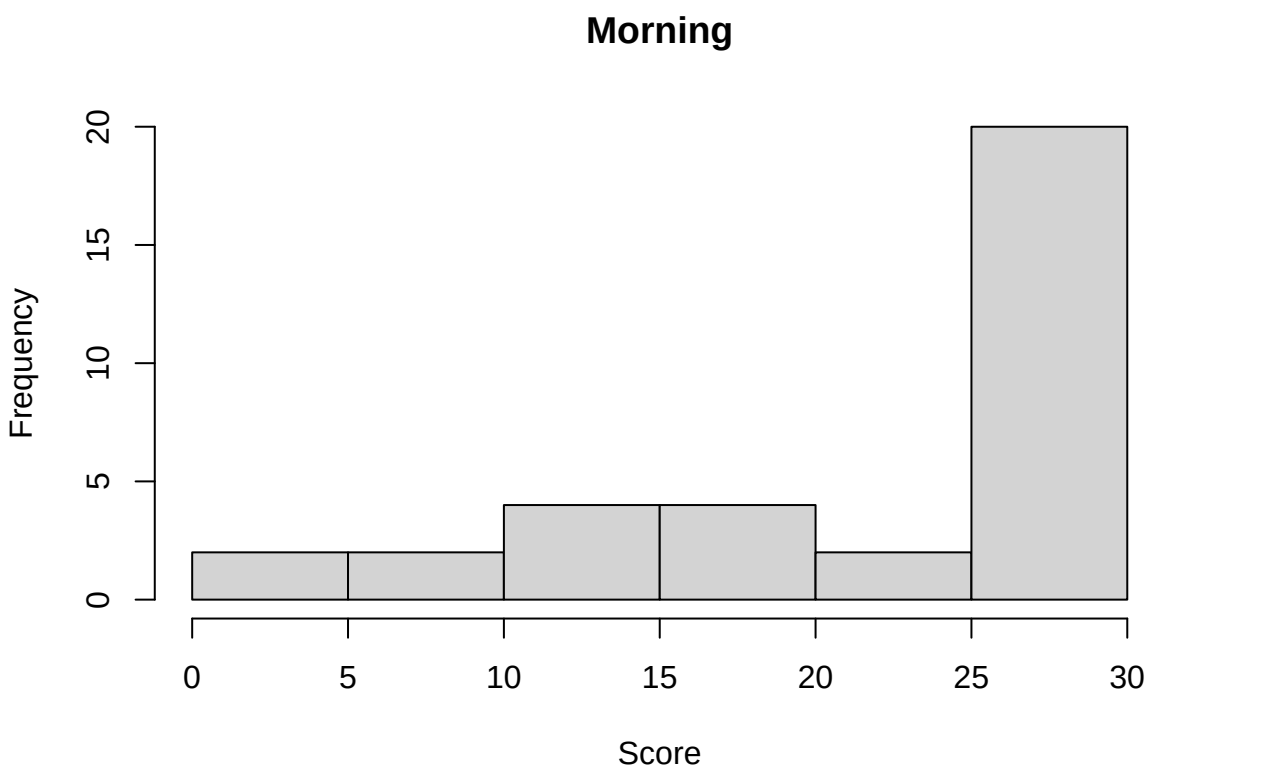
☐ Have H0 and HA been formulated?

☐ Are H0 and HA mutually exclusive and complete? (such that either one or the other can be true, but not both or neither)

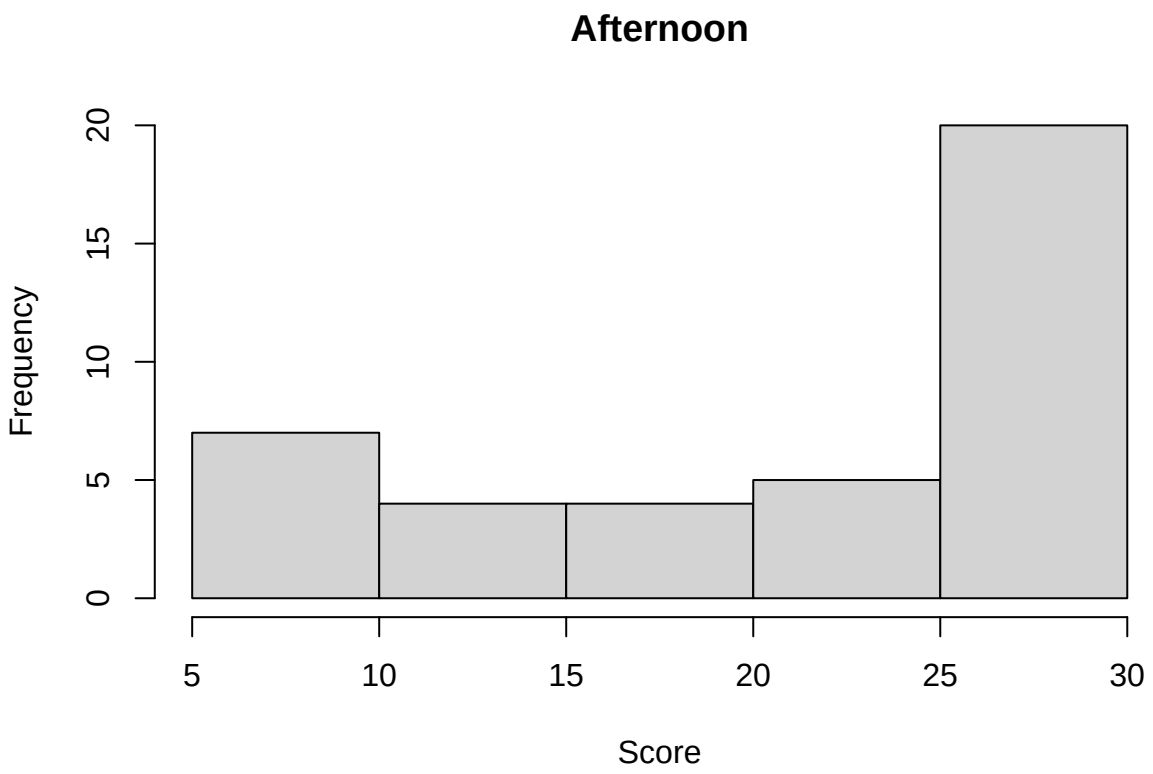
☐ Has a suitable test been identified and has the choice of test been explained? Full points can only be given if there is an explanation!!

☐ What test is being used? It is tempting here to want to do a t-test, so this may be your first idea. If you do a t-test, you have to check the assumptions, and the first assumption is normality. Let's have a look:

```
hist(stroop[stroop$Time=="Morning","Score"],main="Morning", xlab="Score")
```




```
hist(stroop[stroop$Time=="Afternoon","Score"],main="Afternoon", xlab="Score")
```



Clearly, we are not dealing with normal distributions. You will not be punished for having had the idea of doing a t-test at first. But you will lose points if you go through with it without testing assumptions **or** if you test the assumption of normality, find that the data is not normal, and then proceed with a t-test anyway!

Full points can only be given if you decide on a different test. Two tests are possible here: A Wilcoxon ranked sum test or a simulation-based test (re-label "Morning" and "Afternoon" 1000 times or so and compute a p-value for that). Either of these tests is a good choice

☐ Has the test been correctly conducted?

Let me show you what I get in both cases. For the non-parametric test I have

```
wilcox.test(Score~Time,stroop, alternative="two.sided")
```

```
## Warning in wilcox.test.default(x = c(7L, 15L, 27L, 30L, 27L, 12L, 30L, 7L, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Score by Time
## W = 599, p-value = 0.3774
## alternative hypothesis: true location shift is not equal to 0
```

Note that in this case, the Wilcoxon test is two-sided, which corresponds to an Alternative Hypothesis that the Morning and Afternoon group are different.

If your HA was that the Morning group has higher scores than the Afternoon group, then you need to use a one-sided test to go with it.

For the simulation-based test, you should have done something like this (NOTE that this is just an example; there are multiple correct ways of doing this!):

```
morning <- stroop[stroop$Time=="Morning", "Score"]
afternoon<- stroop[stroop$Time=="Afternoon", "Score"]
experiment_difference <- abs(mean(morning)-mean(afternoon))

diffs <- {}

for (i in 1:1000) {
  morning_rows <- sample(74,34,FALSE)
  new_morning <- stroop[morning_rows,"Score"]
  new_afternoon <- stroop[-morning_rows,"Score"]
  diff <- abs(mean(new_morning)-mean(new_afternoon))
  diffs <- c(diffs,diff)
}

p <- sum(diffs >= experiment_difference)/1000
p
```

```
## [1] 0.52
```

☐ Has **exactly** one test been run? While it is ok to do *either* a Wilcoxon test or a simulation-based test, you cannot do *both*. (If they gave you different results, how would you decide?!) You will lose points if you do both

☐ Is the result of the test correctly interpreted? This means two things: First, the p-value has been compared to 0.05: Here, the p-value is larger than 0.05, and therefore we cannot reject the Null Hypothesis. Second, we need an interpretation in terms of the original biological research question! We did not find a difference in performance on the Stroop test according to time of day (Morning or Afternoon).

Name one way in which the study could be improved or followed up on. [5 points]

☐ Has a suggestion been made for improvement or follow-up?

☐ Is the suggestion concrete? For instance, you can't just say "Do better experiments!" Also, you can't just say "Gather more data points in both groups". Because why do you think there aren't enough data points? How many data points would you need? Unless you can give an explanation for that, "collect more data" is just a lazy suggestion.

2. Marathon finishing times [25 points in total]

Import the dataset, and plot it in a way that addresses the question we are interested in. [10 points]

Check for the following:

☐ Has the dataset been imported? Again, we don't need to see the code here, but you should have done something like this:

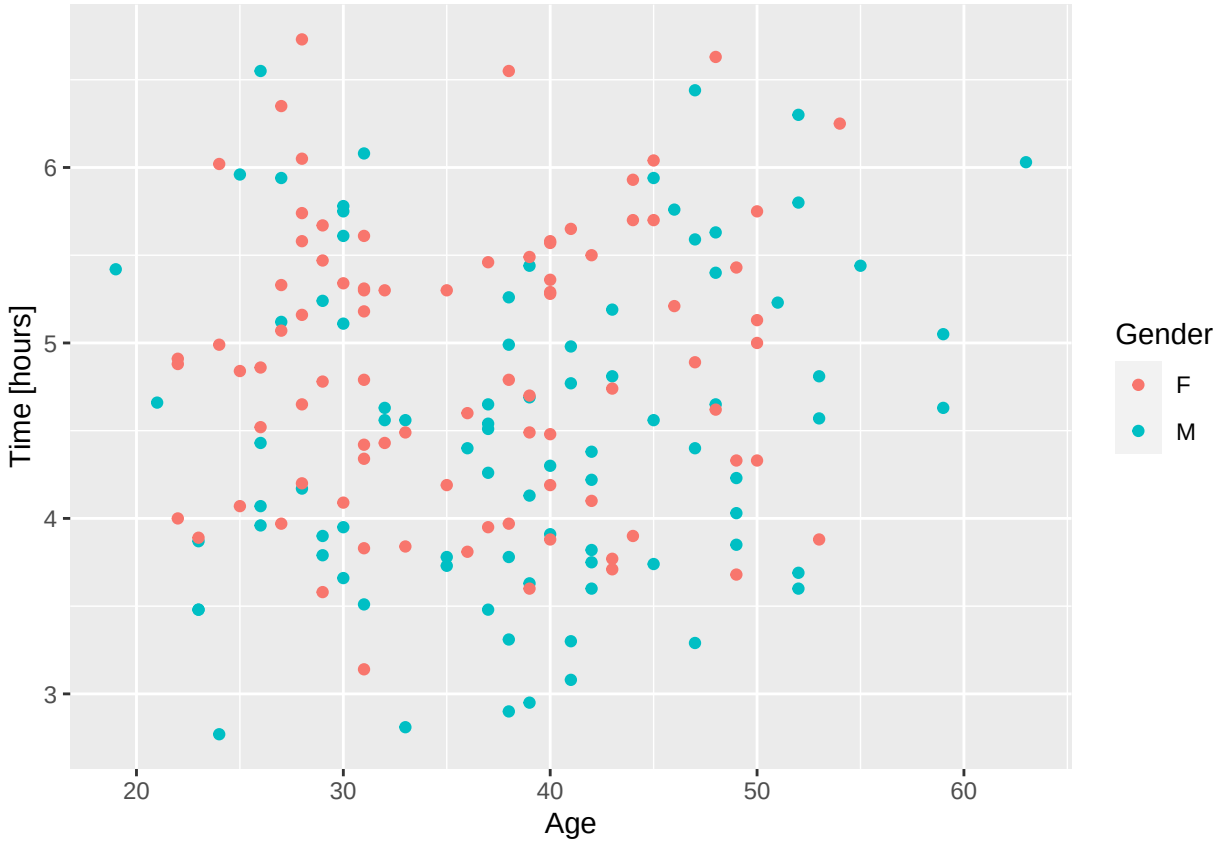
```
marathon <- read.csv("Chicago2013_random_finishers.csv")
head(marathon)
```

```
##                  Name Country Gender Age Time
## 1 Benjamin Reifenberg     USA      M  24 2.77
## 2   Greg Humrichouser     USA      M  33 2.81
```

```
## 3        Stephen Peck     CAN      M  38 2.90
## 4     Rodrigo Llaguno     MEX      M  39 2.95
## 5       Octavio Hoyos     MEX      M  41 3.08
## 6  Kenneth Mcwilliams     USA      M  47 3.29
```

☐ Has the data been plotted in a useful way? Useful here means we need to see the effect of age *and* gender on finishing time. For instance, this would work:

```
q <- ggplot(marathon, aes(x=Age,y=Time, col=Gender))
q <- q+xlab("Age") + ylab("Time [hours]")
q <- q + geom_point()
q
```



Other plots could be fine as well, as long as they show the effects of Gender and Age

☐ Axes are correctly labelled, colours are clear.

☐ The plot shows the data, and only the data. In particular, don't show regression lines (they constitute an interpretation of the data, and unless you have done a statistical test for regression, you don't know how good an interpretation it is.)

What are the average finishing times and standard deviation for each gender? What are the average finishing times and standard deviation for each age quartile? [10 points]

This just requires a bit of grouping. For gender:

```
men <- marathon[marathon$Gender=="M","Time"]
women <- marathon[marathon$Gender=="F","Time"]
mean(men)
```

```
## [1] 4.517882
```

```
sd(men)
```

```
## [1] 0.9243184
```

```
mean(women)
```

```
## [1] 4.883765
```

```
sd(women)
```

```
## [1] 0.8073826
```

By age quartile:

```
q25 <- quantile(marathon$Age,0.25)
q50 <- mean(marathon$Age)
q75 <- quantile(marathon$Age,0.75)

youngest <- marathon[marathon$Age <= q25, "Time"]
young_middle <- marathon[(marathon$Age > q25 & marathon$Age <= q50), "Time"]
old_middle <- marathon[(marathon$Age > q50 & marathon$Age <= q75), "Time"]
oldest <- marathon[marathon$Age > q75, "Time"]
```

(Should it be "smaller than or equal" or just "smaller than"? Honestly, it doesn't matter, as long as some attempt was made to compute score by quartiles.)

Youngest:

```
mean(youngest)
```

```
## [1] 4.8282
```

```
sd(youngest)
```

```
## [1] 0.9139814
```

Younger middle:

```
mean(young_middle)
```

```
## [1] 4.448182
```

```
sd(young_middle)
```

```
## [1] 0.7383193
```

Older middle:

```
mean(old_middle)
```

```
## [1] 4.451818
```

```
sd(old_middle)
```

```
## [1] 0.8481561
```

Oldest:

```r
mean(oldest)
```

```
## [1] 5.001395
```

```r
sd(oldest)
```

```
## [1] 0.8914023
```

Full points should be given if all of those numbers have been correctly computed and reported.

If you had to suggest a statistical test to determine the effect of age quartile and gender on marathon finishig time, what test would you suggest, and why? [5 points]

☐ An appropriate test is suggested, **and** the selection of test is explained. For instance here, because we are looking at two factors (gender and age group), a two-way ANOVA would be a good choice (provided assumptions are met!) If you suggest an ANOVA, you should also explain whether it should be with or without interactions, and why. ☐ Suggesting the test and explaining the suggestion is enough. Note that we don't ask you to run the test. You will not get additional points for running the test! All it makes you do is lose time!

3. Antiviral drug [25 points total]

What are the Null Hypothesis and the Alternative Hypothesis in Prof. Liu's trial? [6 points]

☐ H0 and HA have been formulated, for instance:

- H0: Recovery time is the same independent of treatment
- HA: Reocvery time differs according to what treatment has been administered.

☐ H0 and HA should be clear and precise. For instance, you can't just say:

- H0: There is no effect
- HA: There is an effect

☐ H0 and HA have to be formulated so that exactly one of them has to be true. A commonly-made error is this:

- H0: There is no difference between treatments.
- HA: The new treatment works better than the current treatment.

What is the problem here? Well, what if the new treatment works *worse* than the current treatment? Then H0 is false **and** HA is also false. That can't be happening. The universe will explode! Please take care of the universe! (Actually, the universe is OK. But hypothesis testing won't work, because it relies on the fact that either H0 or HA is true.)

For her statistical analysis, Prof. Liu uses the commonly used significance level α of 0.05. What type of error does this relate to, and how? [10 points]

☐ There is a precise explanation, along these lines: If H0 is true (there is no effect), then the p-value is the probability of seeing data as or more extreme than the one we saw in our experiment. We reject H0 if p is smaller than α. That means that we accept a probability of up to α of mistakenly rejecting H0, even if it's true. This is exactly the probability of making a Type 1 error.

After ensuring that the sample size is big enough and that the assumptions of the statistical tests are met, Prof. Liu runs a statistical test and gets a p-value of 0.059. What is a p-value? Based on this result, what should Prof. Liu report as the outcome of this study? [9 points]

☐ A p-value is correctly defined: It is the probability of seeing data as or more extreme as the data seen in the experiment *if* H0 is true. Go through your words very carefull!! A p-value is **NOT** the probability that H0 is true. It is **NOT** the probability of the data.

☐ It is correctly identified that $0.059 > 0.05$ and therefore H0 cannot be rejected.

☐ An interpretation is given in terms of the original biomedical problem, in this case: We cannot conclude that the new antiviral treatment is any different from the already available treatment.

☐ There is no trying to turn this into a positive result, for instance by saying that 0.059 is quite close to 0.05 so it still kind of counts. (It doesn't. That's what α is for, it's a decision cutoff.) Similarly you can't say that the only reason it's not significant is that there isn't enough data, because the instructions clearly say that Prof. Liu has already ensured that the sample is big enough. Finally, you can't say that the result is "trending towards significance", "approaching significance", "very nearly almost significant" or anything similar. Significance testing is a yes/no thing. (We will look at other types of statistics that go beyond this binary decision in semester 2, when we look at Bayesian statistics.)

Overall presentation and R Markdown (25 points in total)

For this last part, check the following

☐ Was R Markdown used to knit a pdf file? It's fine if you managed to knit to a Word file and then converted that to a pdf. Less fine if you produced an R Markdown file, but did not manage to knit it. Even less fine if you did not produce an R Markdown file.

☐ Is all code provided? The only exception we make are file import commands if the file path would reveal your identity.

☐ Is all code provided *completely* ? Check especially for lines of code that are too long to show up in the knit, like this one:

```
ggplot(hamster, aes(x=lifespan, fill=group)) + facet_grid(cols=vars(group)) + geom_histogram(binwidth =
```

This is a problem, because it means that the code is no longer reproducible to the reader of the knitted document. For full points, longer lines of code should be broken down, for instance like this:

```
p <- ggplot(hamster, aes(x=lifespan, fill=group))
p <- p+ facet_grid(cols=vars(group))
p <- p+ geom_histogram(binwidth = 5)
p <- p + xlab("life span (weeks)")
p <- p + scale_fill_manual(values=c("dodgerblue", "tomato1"))
p
```

☐ Is the code reproducible? To assess this, markers will select one of the three questions at random and try to reproduce your results. Note that here, we don't care if you chose the right analysis approach, or if your code is correct; all we care about is whether it's reproducible.

☐ Is the file nicely formatted? It doesn't have to be a work of art, but it's nice if there are section headers for each of the questions (like there are in this document), so that it's clear where one question ends and the next begins.

```
# Also, the advantage of R Markdown it that allows for text,
# code, and results to nicely coexist in a file. Therefore,
# longer chunks of text (such as explanations) should be
# written as actual text, not as a comment in a code chunk
# (like this one!) - Actual text outside of code chunks is
# much easier to read.
```

☐ Is the writing clear? We don't care so much about smaller typos or perfect English. But the writing should be clear and understandable.