# ADS2 Coding Challenge 1: Submission

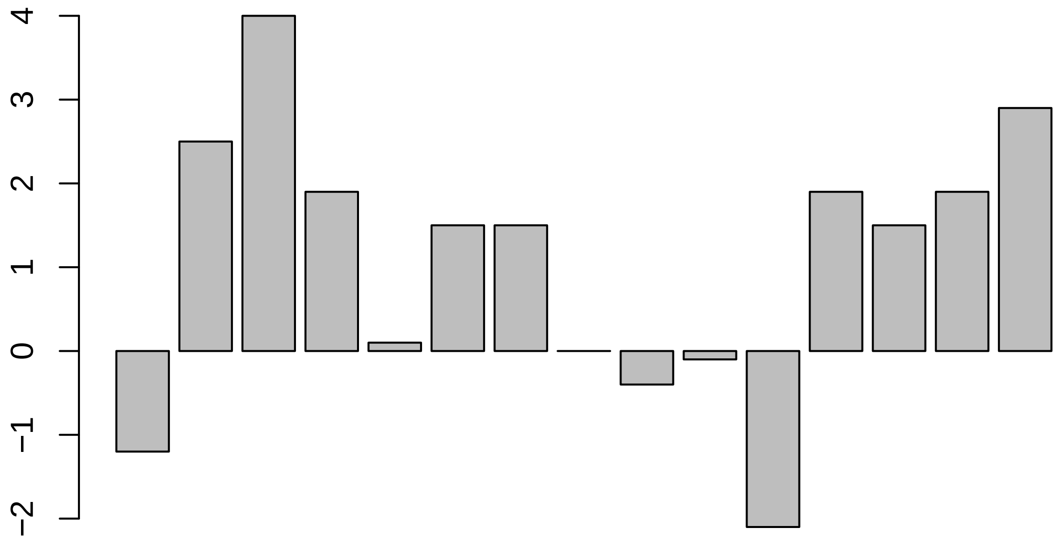
*Roll number: 838*

*2020-01-08 15:58:40*

## 1. Side effects of medication

### Import the dataset and visualise the data in a useful way

```
barplot(weight$weight_changes)
```



### Is there weight gain in mice that have been treated with this medication? Choose and conduct an appropriate test. Explain why the test is appropriate, and discuss what the results mean

```
t.test(weight$weight_changes)
```

```
##
##  One Sample t-test
```

```
##
## data:  weight$weight_changes
## t = 2.5112, df = 14, p-value = 0.02492
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1546779 1.9653221
## sample estimates:
## mean of x
##      1.06
```

Explanation: one-sample t-test is used because sample size is small (n<30). The results shows a p-value smaller than 0.05, by which we cannot reject the null hypothesis that there is no difference in weight gain after treating medication. Therefore, we can say with 95% confident that there is weight gain in mice that have been treated with this medication. ## Name and explain one way in which the experiment could be improved or one possible direction for future study. The sample size (n = 15) is small. To convincely make a further conclusion, larger sample size is needed.

## 2. Drug use among college students

### What proportion of students are likely to have used illegal drugs?

```
x = 2*(108/250-0.5*182/365)
x
```

```
## [1] 0.3653699
```

### What is the reason for setting up the survey in such a complicated way?

When getting information about sensitive topics, interviewers might hide the truth, which will cause sampling bias. This complicated way of setting up a survey can avoid sampling bias in a certain degree.

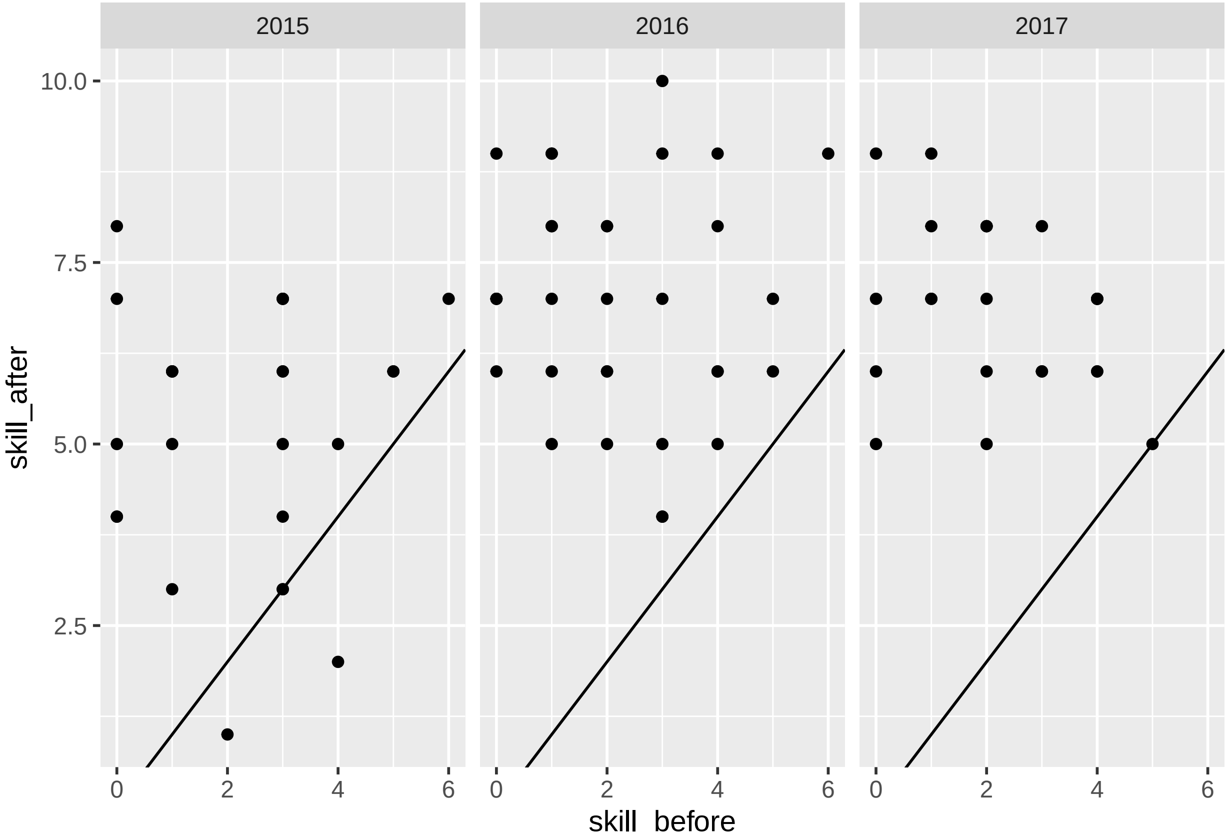
# 3. Student improvement in a beginning programming course

### Import the dataset and visualise the dataset in a useful way

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(data = student, aes(x = skill_before, y = skill_after, group = year))+
  geom_point()+
  geom_abline(slope = 1, intercept = 0)+
  facet_wrap(.~year)
```



## Explain in a few short sentences why this is a useful way to look at the data The plot is separated in three plots, showing the group of different years. Every dot represent a student and his/her skill before and after can be read in x and y axes. An ectra line is added to represent there is no difference before and after. Points above the line shows increased programming skill after taking the course, and vice versa. ## You may have noticed that for some students, taking the course *decreases* their programming skills. For how many students is this true?

```
nrow(student[which(student$skill_after-student$skill_before<0),])
```

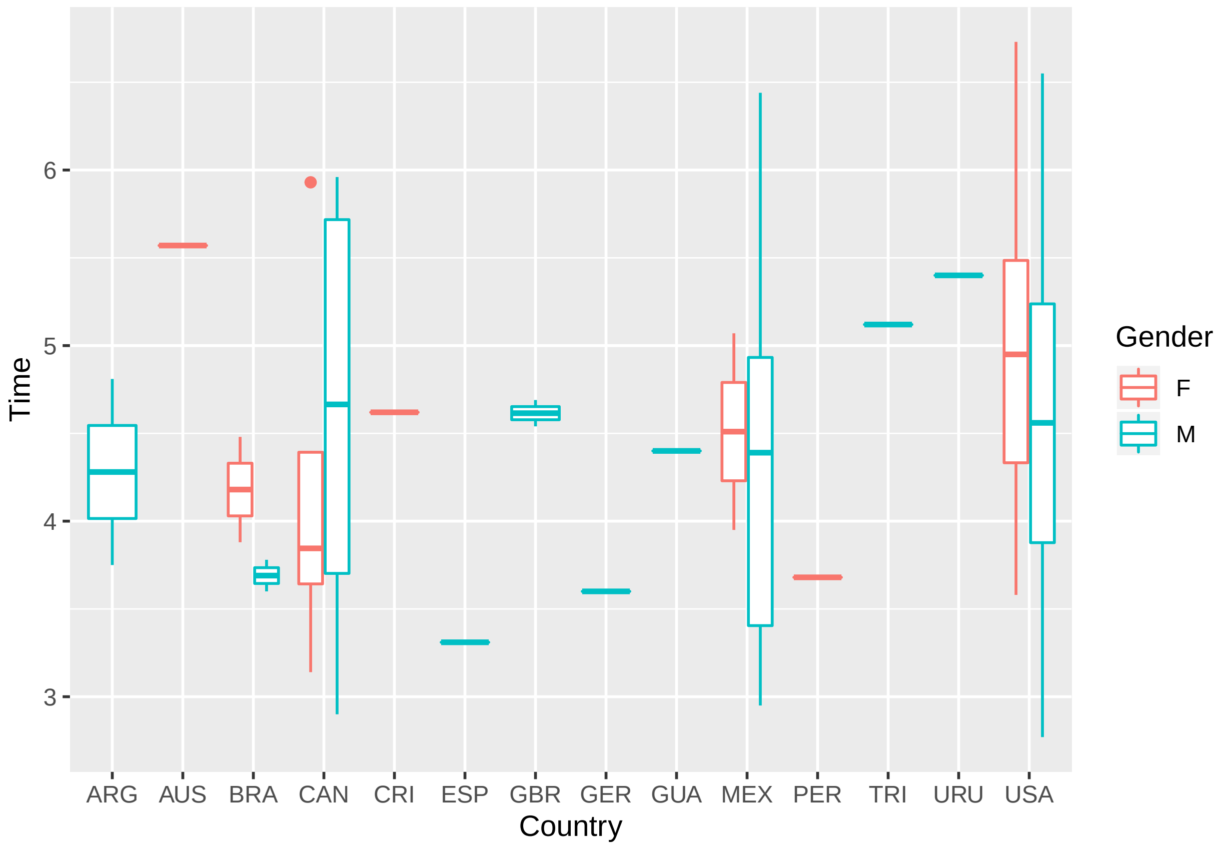
```
## [1] 2
```

### Why do you think this is the case, and how might it be fixed?

In the plot above, we can see that there are two points below the line, indicating that there are two students who have *decreased* their programming skills after taking the course. If there is no measuring mistakes, the data should not be fixed.

## 4. Finishing times in the Chicago Marathon

### Import the dataset and visualise the data in a useful way



### Is the average difference between two runners in the same group, smaller, bigger, or equal to the average difference between two runners in different groups? What does this tell you?

This probably can be done by the imcomplete code below.

This tests the variance difference between and within groups.

```
America <- chicago[which(chicago$Country == "USA"),]
Not_America <- chicago[which(chicago$Country != "USA"),]
F_America <- America[which(America$Gender == "F"),]
M_America <- America[which(America$Gender == "M"),]
```

```
F_Not_America <- Not_America[which(Not_America$Gender == "F"),]
M_Not_America <- Not_America[which(Not_America$Gender == "M"),]
```

```
same_group <- c()
dif_group <- c()
for (i in 1:1000) {
  index1 <- sample(1:nrow(chicago),"Time")
  s1 <- chicago[index1,]

  trial2 <- chicago[-index1,]
  index2 <- sample(1:nrow(trial2),"Time")
  s2 <- chicago[index2,]

  if (s1$Country == "USA") {
    same_group <- c(same_group,abs(s1$Time-s2$Time))
  }

  if (s1$treatment != "USA") {
    dif_group <- c(dif_group,abs(s1$Time-s2$Time))
  }
}
```

## Is there a statistical test that you could do to determine whether gender and/or country of origin has an effect on finishing time? What assumptions need to be met in order to conduct that test, and are those assumptions met in this dataset?

ANOVA.

Assumptions 1. Independent random sampling. 2. Normality of residuals. 3. Equality of variances.

1. Independent random sampling is met. Female and male runners are randomly selected from the whole sample.
2. The following is the code to test normality of residuals, if p-value is smaller than 0.05, it is a normal distribution

   ```
   chicago$Country <- factor(chicago$Country)
   model <- aov(Gender ~ Time * Country, data = chicago)
   shapiro.test(resid(model))
   ```

3. The following is the code to test the equality of variances, if the red line is approximately horizontal, that means it meets the assumption.

   ```
   plot(model,1)
   ```