

# Task 2 - R Studio

## Importing Data

The data we will look at today comes from the Coffee Quality Institute. Over 1000 beans were reviewed by trained reviewers, who considered things like aroma, acidity and balance to come up with a total score for each bean.

Follow these steps:

1. Open R Studio.
2. Either click "File" -> "New File" -> "R Script" or press CTRL+SHIFT+N. The window that appears is where we will type our code.
3. Now either click "Session" -> "Set Working Directory" -> "Choose Directory" or press CTRL + SHIFT + H.
4. Choose the folder where the file "coffee\_adjusted.csv" is located.

The following code will allow you to import the dataset:

```
coffee <- read.csv("coffee_adjusted.csv")  
head(coffee)
```

The first line creates a variable called `coffee` and assigns the csv file to it. The `head` function prints the first 6 rows of the dataset- this is extremely useful for checking you have imported the dataset as expected.

### Info

To run a line of code in R Studio, you must press CTRL+ENTER or click "Run" whilst your cursor is on the line you wish to run.

You can also run multiple lines of code by highlighting all the lines you wish to run and pressing "Run" or CTRL+ENTER.

## Exploring Data in R

Below are examples of R code for generating simple statistics, creating tables and creating graphical displays.

```
### Calculate the mean total score for Brazil  
mean(coffee$Total.Score.Brazil, na.rm = TRUE)  
### Calculate several summary statistics for the total score of Natural, Dry  
beans
```

```

summary(coffee$Total.Score.Natural.Dry)
### Calculate the standard deviation and interquartile range for total Aroma
Scores
sd(coffee$Aroma.Score.All, na.rm = TRUE)
IQR(coffee$Aroma.Score.All, na.rm = TRUE)
### Create a table showing which countries the beans came from
table(coffee$Location.All)
### The same table as a percentage table
prop.table(table(coffee$Location.All))*100
### Create a barplot showing the distribution of processing methods
barplot(table(coffee$Processing.Method.All),
        main="Processing Methods",
        xlab="Methods",
        ylab="Frequency",
        col=c("coral1", "coral2", "coral3", "yellow1", "yellow2", "yellow3"))
### Create a scatterplot with Acidity Scores on Balance Scores
plot(coffee$Balance.Score.All, coffee$Acidity.Score.All,
     main="Acidity Scores on Balance Scores",
     xlab="Balance Scores",
     ylab="Acidity Scores",
     pch=19)
### Create boxplots comparing the total scores for beans for Brazil and the US
boxplot(coffee$Total.Score.Brazil, coffee$Total.Score.US,
        main="Total Scores for Brazil and the US",
        xlab="Country",
        ylab="Total Score",
        names=c("Brazil", "US"),
        col = c("yellow", "red"))
### List all colours
colours()

```

## Task

1. Copy each line of code from the examples above into R Studio and run the code. Notice what happens when a `$` is typed. You can use the arrow keys and the enter key to code with speed.
2. Find the mean total score for all the beans (`Total.Score.All`). Find the mean total score for China, how does this compare to all the beans?
3. Produce summary statistics for the total score of washed/wet beans. Also find the standard deviation and the interquartile range.
4. Create a table showing the types of processing methods. Also produce a similar table showing the percentage of each type represented in the dataset.
5. Create a barplot showing the percentage of each processing method. Hint: use your previous answer.
6. Create a scatterplot of Aftertaste Scores on Acidity Scores.
7. Create boxplots comparing the total scores for beans from the US and China.

8. R can be used to do simple computations, like a calculator. For example you can run things like `2+3*10` and `mean(c(2,4,6))/10`. Calculate the semi-interquartile range for the total score for all the beans.
9. The distribution of numerical data can be seen using for example `hist(coffee$Total.Score.Washed.Wet)`. For the numerical data in the dataset, determine whether the mean or the median would be the most appropriate measure of central location.
10. Find another dataset online. Investigate.