# Exercise sheet: Day 2

***Vangelis Theodorakis, Fatemeh Behjati, Julien Gagneur, Marcel Schulz***

**12 April, 2021**

## Contents

# 1 Setup

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
library(readxl)
library(dplyr)
```

# 2 Introduction to ggplot

The `iris` data is included in the ggplot2 package. First load `ggplot2` package, then load `iris` data by `data(iris)`. Check `iris` data with `head(iris)`.

1) Are there any relationships/correlations between petal length and width? How would you show it?
2) Do petal lengths and widths correlate in every species?

# 3 Histograms

Get the *gtex-annotation.csv* data and do a histogram of the RIN (RNA integrity number) column using 10, 20, 50, 100 bins to see how this affects the visualisation.

# 4 Boxplots

Get the *gtex-annotation.csv* data and do some boxplots of the RIN (RNA integrity number) column against the age groups. Do you see something interesting? Do the same using violin plots. Now try to combine the violin and the boxplot into one plot (use `width = 0.2`).

# 5 Scatterplot

Make a scatterplot between the fake.age and the RIN for heart. Do you see any associasion between fake.age and RIN? Now color the points by `sex` so that you have the labels `Male` and `Female` on the legend. Do you see any associasion between fake.age and RIN when it is controled for sex?

# 6 Understanding a messy dataset

The following file describes the number of times a person bought a product "a" and "b"

```
messy_file <- file.path('extdata', 'example_product_data.csv')
messy_dt <- fread(messy_file)
messy_dt
##            name producta productb
## 1:     John Doe       NA       12
```

```
## 2:    Marry Doe        3        1
## 3: John Johnson        5        1
```

Why is this data-set messy? Which columns should a tidy version of this table have?

# 7 Fixing a messy dataset

Read the weather dataset `weather.txt`. It contains the minimal and maximal temperature on a certain city (id) over different dates (year, month, d1-d31). Why is this dataset messy? How would a tidy version of it look like? Create its tidy version.