Auckland University of Technology
Department of Mathematical Sciences
COMP824: Week 5 Lab

Some of the exercises in this lab are based on those found in Wickham and Grolemund, R for Data Science.

Getting started

(1) Install the `tidyverse` package (only required if this is the first time you are using tidyverse on this computer).

Uncomment this following line (i.e. remove the `#`) and change the chunk option to `eval=FALSE`. Knit your file and see what happens.

```
#install.packages("tidyverse")
```

(2) Load the `tidyverse` package

```
library(tidyverse)
```

Tip

You can get help in R by typing "?" followed by the function name. For example ?ggplot

R Markdown

(3) You can change what gets displayed when you compile your Rmd file by changing the chunk options.

Some examples are shown below. Look at the code in the Rmd file and output in the PDF file and figure out what each one does.

- `echo = FALSE`
- `echo = TRUE`

```
x <- 1:10
x
```

- `results = "hide"`

```
x <- 1:10
x
```

- `results = "hold"`

```
x <- 1:10
x
y <- 2:11
y
```

- `include = TRUE` (change to FALSE)

```
z <- 1:10
```

(4) Have a look at this page https://yihui.org/knitr/options/#code-chunk to get more information about different chunk options.

(5) You can add code within an R Markdown paragraph. For example, the mean of `z` is 5.5 (inspect within Rmd file).

dplyr

(6) Load and inspect the `flights` dataset

```
library(nycflights13)
flights
```

(7) What does the variable "air_time" represent?

(8) Create a dataset called `march1` containing flights which departed on 1st March 2013. How many flights departed on 1st March 2013?

(9) For the flights which departed on 1st March 2013, create a new variable called `max_delay` which contains the maximum of the departure delay and the arrival delay. Arrange in descending order by the new variable so the first row contains the flight with the largest departure or arrival delay. Which flight had the greatest delay and how long was the delay?

Hint: You will need to use the command `rowwise()` (see PDF file)

```
march1 %>% rowwise() %>% mutate(...)
```

(10) For the flights which departed on 1st March 2013, what was the destination of the one with the biggest departure delay?

(11) Which destination has the greatest mean departure delay in 2013? Hint: use `group_by` and `summarise`

(12) Use summarise to determine the number of flights by each carrier.

Hint: `summarise(n=n())`

(13) For flights which departed on 1st March, create a boxplot showing the departure delay, with one box for each "carrier". Which airline seems to have the worst on-time performance?

tibble

(14) Consider the iris data set.

- Run the following code and inspect the difference in output.

```
iris
(iris_tb <- as_tibble(iris))
```

- Using `iris_tb`, create a ggplot to compare `Sepal.Width` and `Sepal.Length`, with a different colour for each `Species`. Write 2 - 3 sentences describing your plot.

(15) Create the following tibble.

```
set.seed(123414)
annoying <- tibble(
  `1` = 1:10,
  `2` = `1` * 2 + rnorm(length(`1`))
)
```

Practice referring to non-syntactic names in the following data frame by:

- Extracting the variable called 1. The desired output is shown below.
- Plotting a scatterplot of 1 vs 2.
- Creating a new column called 3 which is 2 divided by 1. The desired output is shown below.
- Renaming the columns to one, two and three. The desired output is shown below.

Application

Pick one of the following datasets:

- `ggplot2::mpg`
- `nycflights13::flights`
- `ggplot2::diamonds`

(16) Look at the help page in R for your chosen dataset, e.g. `?flights`.

(17) Write 3 questions about this dataset.

e.g. How many flights were late on Christmas day?

(18) Transform the data and/or create an summary statistics to investigate each of your questions.

Further practice

(19) Read R4DS chapters 5, 6, 10 and do the exercises.