Minilab 1b Worksheet

Introduction to the "tidyverse"

In this minilab, we will get a taster of the "tidyverse" collection of R packages and construct some plots using ggplot2.

1. Why do we learn R in this module?

Python and R are the "dynamic duo" (like Batman and Robin) of Data Science.

- In Python, there are a number of very useful libraries for Data Science including numpy, pandas and matplotlib (these are part of the scipy ecosystem).
- In R, there are a number of very useful packages for Data Science including dplyr and ggplot2 (these are part of the tidyverse).

Since you already know some Python, it will give you some additional skills if you also learn some R (especially the tidyverse). R is very strong in statistics since it has been developed by statisticians for statisticians. The tidyverse makes R easy to use for data science.

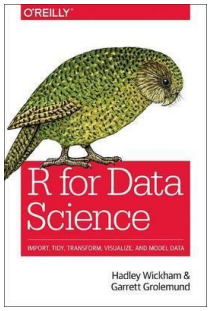
Interesting reading: https://www.datanami.com/2020/08/11/r-and-python-the-data-science-dynamic-duo/

2. Introduction to the "tidyverse" for Data Science

We will spend a good amount of time in this module discussing data visualisation. It serves many roles in data analysis. We use it to gain understanding of dataset characteristics throughout analyses and it is a key element of communicating insights we have derived from data analyses with our target audience.

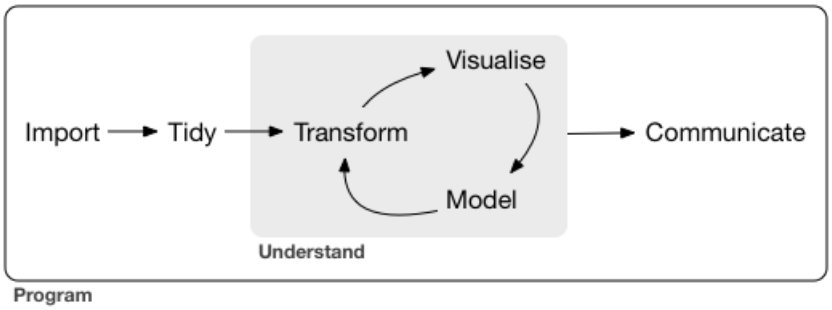
The "tidyverse" is a collection of R packages designed for Data Science. The design and use of these packages (including lots of examples) are described very well in the book:

Wickham, H. and Grolemund, G. (2017) *R for Data Science*, O'Reilly.



The book is available **free** as a website at https://r4ds.had.co.nz/.

A typical data science project looks something like the following. Each of the R packages in the tidyverse addresses some part of this process.



The R package in the tidyverse that handles creating beautiful graphics is ggplot2. It creates plots using a "grammar of graphics" approach.

(1) To download and install the complete tidyverse, type the following into the RStudio console pane (bottom left of the RStudio IDE). *Note the quotes.*

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

If you are asked for a CRAN mirror, choose https://cloud.r-project.org/. There are quite a few packages to install so this might take several minutes to complete. *CRAN is the Comprehensive R Archive Network.*

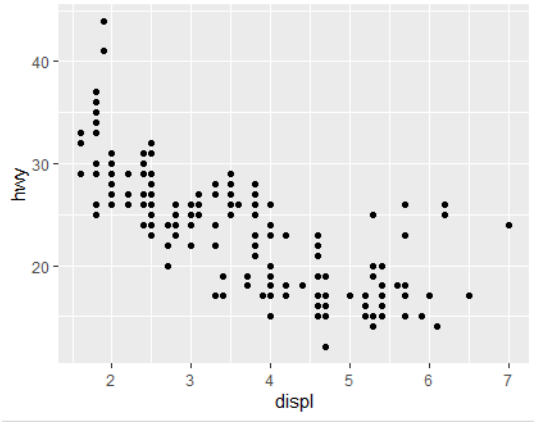
(2) To use the tidyverse packages, you need to load them using the following "library" command (note the **lack of quotes** around tidyverse this time) so that the packages are available for the rest of your R session. You will need this line at the top of any R script you write that produces graphics using ggplot2.

```
library(tidyverse)
```

(3) The dataset "mpg" comes as part of the "ggplot2" package (so we don't need to load it into R separately). It contains data that compares the fuel economy data from 1999 and 2008 for 38 popular models of car. Two of the variables in mpg are `displ` (car engine size in litres) and `hwy` (car fuel economy on the highway in miles per gallon).

Create a new R script (File | New File | R Script), enter the following R code and save your file in an appropriate place. Select the code you want to run (use ctrl-a to select all the code in the editor) and click on the Run button to run it in the R console. This will plot a *scatterplot* with displ (engine size) on the horizontal axis and hwy (fuel economy) on the vertical axis. Each dot represents one row of the mpg data table.

```
library(tidyverse)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=hwy))
```



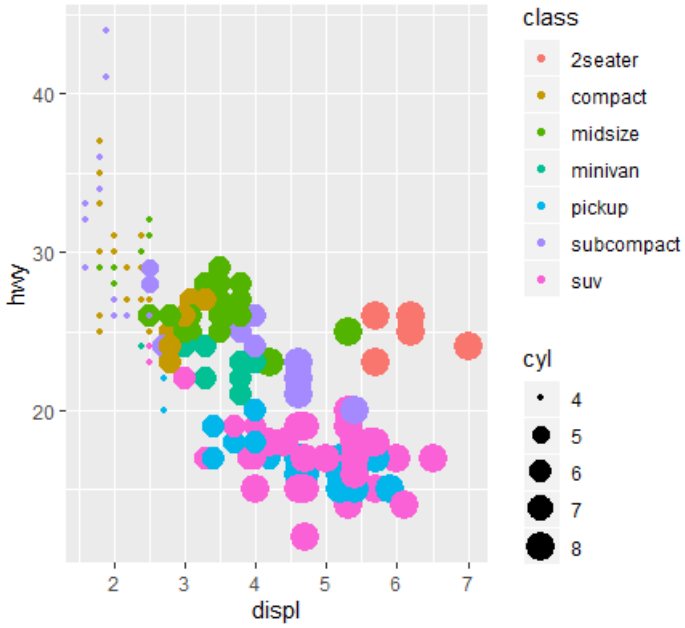
The R package ggplot2 is designed to produce beautiful graphics. We specify:

1. The **data** that goes into a plot, i.e., a data table of entities (rows) and attributes (columns). In this case it is the mpg data table.
2. The **mapping** between data attributes and graphical (aesthetic) characteristics (note the "aes" function). In this case it is a coordinate system in which displ is mapped to the horizontal axis (x-axis) and hwy is mapped to the vertical axis (y-axis). We could also map an attribute to the colour, shape or size of the plotting symbol. *It is ok to use "color" (US spelling) or "colour" (UK spelling) in ggplot2.*
3. The **geometric** representation of these graphical characteristics. In this case it is that the plotting symbol is a point (note the "geom_point" function).

We will look at building graphical plots in more detail in subsequent minilabs.

(4) Modify your R code to use cyl (number of cylinders) to select the size of the plotting symbol and class to select a colour (a *bubbleplot*). Legends that explain the levels of colour and sizes is automatically added to the plot. For more details about the mpg dataset, see https://ggplot2.tidyverse.org/reference/mpg.html.

```
library(tidyverse)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=hwy, colour=class, size=cyl))
```



Exercise. Note down any patterns or trends you "see" in this plot. Write a sentence and share it on Microsoft Teams chat or on Aula.

Summary

In this minilab, we have installed the tidyverse set of R packages and seen how to create a basic scatterplot using ggplot2. In future minilabs, we will explore the various tidyverse packages used to import, tidy, transform, visualise, model and communicate. We will also use R to undertake exploratory data analysis, build statistical models, and understand data science concepts.

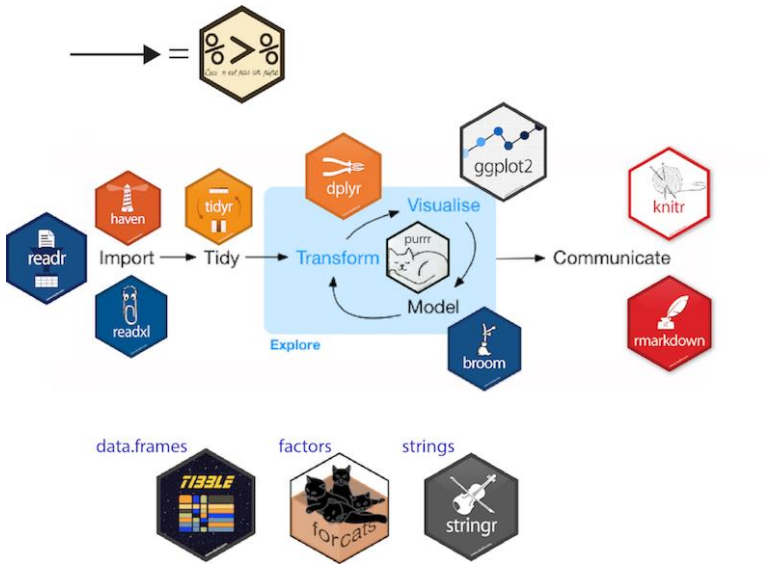


Image from: https://rworkshop.uni.lu/lectures/lecture07_plotting.html#4