Practical session materials | granolarr

Stefano De Sabbata

2021-10-03

Contents

Preface

Stefano De Sabbata

This book contains the *practical sessions* component of granolarr, a repository of reproducible materials to teach geographic information and data science in R. Part of the materials are derived from the practical sessions for the module GY7702 Practical Programming in R of the MSc in Geographic Information Science at the School of Geography, Geology, and the Environment of the University of Leicester, by Dr Stefano De Sabbata.

This book was created using R, RStudio, RMarkdown, Bookdown, and GitHub.

Session info

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04 LTS
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/openblas-openmp/libopenblasp-r0.3.8.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
```

```
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.2  magrittr_1.5    bookdown_0.20   htmltools_0.5.0
##  [5] tools_4.0.2     yaml_2.2.1      stringi_1.4.6   rmarkdown_2.3
##  [9] knitr_1.29      stringr_1.4.0   digest_0.6.25   xfun_0.16
## [13] rlang_0.4.7     evaluate_0.14
```

Chapter 1

Introduction to R

Stefano De Sabbata

1.1 The R programming language

As mentioned in Lecture 1, **R** was created in 1992 by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R is a free, open-source implementation of the **S** statistical programming language initially created at the Bell Labs. At its core, R is a functional programming language (its main functionalities revolve around defining and executing functions). However it now supports, and it is commonly used as an imperative (focused on instructions on variables and programming control structures) and object-oriented (involving complex object structures) programming language.

In simple terms, nowadays, programming in R mostly focuses on devising a series of instructions to execute a task – most commonly, loading and analysing a dataset.

As such, R can be used to program by creating sequences of **instructions** involving **variables** – which are named entities that can store values. That will be the main topic of this practical session. Instructions can include control flow structures, such as decision points (*if/else*) and loops, which will be the topic of the next practical session. Instructions can also be grouped in **functions**, which we will also see in the next practical session.

R is **interpreted**, not compiled. Which means that an R interpreter (if you are using R Studio, the R interpreter is simply hidden in the backend and R Studio is the frontend that allows you to interact with the interpreter) receives an instruction you write in R, interprets and executes them. Other programming

languages require their code to be compiled in an executable to be executed on a computer.

1.1.1 Using RStudio

As you open RStudio or RStudio Server, the interface is divided into two main sections. On the left side, you find the *Console* – as well as the R script editor, when a script is being edited. The *Console* in an input/output window into the R interpreter, where instructions can be typed, and the computed output is shown.

For instance, if you type in the *Console*

```
1 + 1
```

the R interpreter understands that as an instruction to sum one to one, and produces the result (as the materials for this module are created in RMarkdown, the output of the computation is always preceded by '##').

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

Note how the output value `2` is preceded by `[1]`, which indicates that the output is constituted by only one element. If the output is constituted by more than one element, as the list of numbers below, each row of the output is preceded by the index of the first element of the output.

```
##  [1]   1   4   9  16  25  36  49  64  81 100 121 144 169 196 225 256 289 324 361
## [20] 400
```

On the right side, you find two groups of panels. On the top-right, the main element is the *Environment* panel, which is a representation of the current state of the interpreter's memory, and as such, it shows all the stored variables, datasets, and functions. On the bottom-right, you find the *Files* panel, which shows file system (file and folders on your computer or the server), as well as the *Help* panel, which shows you the help pages when required. We will discuss the other panels later on in the practical sessions.

1.2 Interpreting values

When a value is typed in the *Console*, the interpreter simply returns the same value. In the examples below, `2` is a simple numeric value, while `"String value"` is a textual value, which in R is referred to as a *character* value and in programming is also commonly referred to as a *string* (short for *a string of characters*).

Numeric example:

```
2
```

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

Character example:

```
"String value"
```

```
## [1] "String value"
```

Note how character values need to start and end with a single or double quote (' or "), which are not part of the information themselves. The Tidyverse Style Guide suggests always to use the double quote ("), so we will use those in this module.

Anything that follows a # symbol is considered a *comment* and the interpreter ignores it.

```
# hi, I am a comment, please ignore me
```

As mentioned above, the interpreter understands simple operations on numeric values.

```
1 + 1
```

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

There are also a large number of pre-defined functions, e.g., square-root: `sqrt`.

```
sqrt(2)
```

```
## [1] 1.414214
```

Functions are collected and stored in *libraries* (sometimes referred to as *packages*), which contains related functions. Libraries can range anywhere from the `base` library, which includes the `sqrt` function above, to the `rgdal` library, which contains implementations of the GDAL (Geospatial Data Abstraction Library) functionalities for R.

1.3 Variables

A variable can be defined using an **identifier** (e.g., `a_variable`) on the left of an **assignment operator** `<-`, followed by the object to be linked to the identifier, such as a **value** (e.g., `1`) to be assigned on the right. The value of the variable can be tested/invoked by simply specifying the **identifier**.

```
a_variable <- 1
a_variable
```

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

If you type `a_variable <- 1` in the *Console* in RStudio, a new element appears in the *Environment* panel, representing the new variable in the memory. The left part of the entry contains the identifier `a_variable`, and the right part contains the value assigned to the variable `a_variable`, that is `1`.

It is not necessary to provide a value directly. The right part of the assignment can be a **call to a function**. In that case, the function is **executed** on the provided input and **the result is assigned to the variable**.

```
a_variable <- sqrt(4)
a_variable
```

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

Note how if you type `a_variable <- sqrt(4)` in the *Console* in RStudio, the element in the *Environment* panel changes to reflect the new value assigned to the variable `a_variable`, which is now the result of `sqrt(4)`, that is 2.

In the example below, another variable named `another_variable` is created and summed to `a_variable`, saving the result in `sum_of_two_variables`. The square root of that sum is then stored in the variable `square_root_of_sum`.

```
another_variable <- 4
another_variable
```

```
## [1] 4
```

```
sum_of_two_variables <- a_variable + another_variable

square_root_of_sum <- sqrt(sum_of_two_variables)
square_root_of_sum
```

```
## [1] 2.44949
```

1.4 Basic types

1.4.1 Numeric

The *numeric* type represents numbers (both integers and reals).

```
a_number <- 1.41
is.numeric(a_number)
```

```
## [1] TRUE
```

```
is.integer(a_number)
```

```
## [1] FALSE
```

```
is.double(a_number) # i.e., is real
```

```
## [1] TRUE
```

Base numeric operators.

Operator	Meaning	Example	Output
+	Plus	`5+2`	7
-	Minus	`5-2`	3
*	Product	`5*2`	10
/	Division	`5/2`	2.5
%/%	Integer division	`5%/%2`	2
%%	Module	`5%%2`	1
^	Power	`5^2`	25

Some pre-defined functions in R:

```
abs(-2) # Absolute value
```

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

```
ceiling(3.475) # Upper round
```

```
## [1] 4
```

```
floor(3.475) # Lower round
```

```
## [1] 3
```

```
trunc(5.99) # Truncate
```

```
## [1] 5
```

```
log10(100) # Logarithm 10
```

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

```
log(exp(2)) # Natural logarithm and e
```

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

Use simple brackets to specify the order of execution. If not specified the default order is: rise to power first, then multiplication and division, sum and subtraction last.

```
a_number <- 1
(a_number + 2) * 3
```

```
## [1] 9
```

```
a_number + (2 * 3)
```

```
## [1] 7
```

```
a_number + 2 * 3
```

```
## [1] 7
```

The object `NaN` (*Not a Number*) is returned by R when the result of an operation is not a number.

```
0 / 0
```

```
## [1] NaN
```

```
is.nan(0 / 0)
```

```
## [1] TRUE
```

That is not to be confused with the object `NA` (*Not Available*), which is returned for missing data.

1.4.2 Logical

The *logical* type encodes two truth values: True and False.

```
logical_var <- TRUE
is.logical(logical_var)
```

```
## [1] TRUE
```

```
isTRUE(logical_var)
```

```
## [1] TRUE
```

```
as.logical(0) # TRUE if not zero
```

```
## [1] FALSE
```

Basic logic operators

Operator	Meaning	Example	Output
==	Equal	`5==2`	FALSE
!=	Not equal	`5!=2`	TRUE
>	Greater than	`5>2`	TRUE
<	Less than	`5<2`	FALSE
>=	Greater or equal	`5>=2`	TRUE
<=	Less or equal	`5<=2`	FALSE
!	Not	`!TRUE`	FALSE
&	And	`TRUE & FALSE`	FALSE
|	Or	`TRUE | FALSE`	TRUE

1.4.3 Character

The *character* type represents text objects, including single characters and character strings (that is text objects longer than one character, commonly referred to simply as *strings* in computer science).

```
a_string <- "Hello world!"
is.character(a_string)
```

```
## [1] TRUE
```

```
is.numeric(a_string)
```

```
## [1] FALSE
```

```
as.character(2) # type conversion  (a.k.a. casting)
```

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

```
as.numeric("2")
```

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

```
as.numeric("Ciao")
```

```
## Warning: NAs introduced by coercion
```

```
## [1] NA
```

1.5 Tidyverse

As mentioned in the lecture, libraries are collections of functions and/or datasets. Libraries can be installed in R using the function `install.packages` or using `Tool > Install Packages...` in RStudio.

The meta-library Tidyverse contains the following libraries:

- `ggplot2` is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
- `dplyr` provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges.
- `tidyr` provides a set of functions that help you get to tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable.
- `readr` provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes.
- `purrr` enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. Once you master the basic concepts, purrr allows you to replace many for loops with code that is easier to write and more expressive.
- `tibble` is a modern re-imagining of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles

are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code.

- `stringr` provides a cohesive set of functions designed to make working with strings as easy as possible. It is built on top of stringi, which uses the ICU C library to provide fast, correct implementations of common string manipulations.
- `forcats` provides a suite of useful tools that solve common problems with factors. R uses factors to handle categorical variables, variables that have a fixed and known set of possible values.

A library can be loaded using the function `library`, as shown below (note the name of the library is not quoted). Once a library is installed on a computer, you don't need to install it again, but every script needs to load all the library that it uses. Once a library is loaded, all its functions can be used.

Important: it is always necessary to load the `tidyverse` meta-library if you want to use the `stringr` functions or the pipe operator `%>%`.

```
library(tidyverse)
```

1.5.1 stringr

The code below presents the same examples used in the lecture session to demonstrate the use of `stringr` functions.

```
str_length("Leicester")
```

```
## [1] 9
```

```
str_detect("Leicester", "e")
```

```
## [1] TRUE
```

```
str_replace_all("Leicester", "e", "x")
```

```
## [1] "Lxicxstxr"
```

1.5.2 The pipe operator

The pipe operator is useful to outline more complex operations, step by step (see also R for Data Science, Chapter 18). The pipe operator `%>%`

- takes the result from one function
- and passes it to the next function
- as the **first argument**
- that doesn't need to be included in the code anymore

The code below shows a simple example. The number `2` is taken as input for the first pipe that passes it on as the first argument to the function `sqrt`. The

output value `1.41` is then taken as input for the second pipe, that passes it on as the first argument to the function `trunc`. The final output `1` is finally returned.

```
2 %>%
  sqrt() %>%
  trunc()
```

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

The image below graphically illustrates how the pipe operator works, compared to the same procedure executed using two temporary variables that are used to store temporary values.

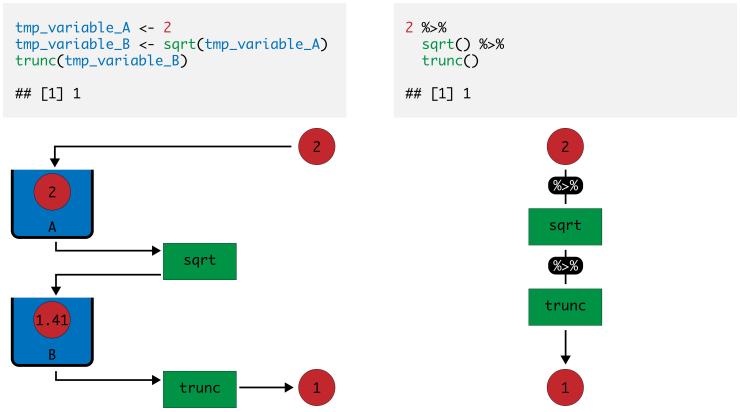


Figure 1.1: Illustration of how the pipe operator works

```
sqrt(2) %>%
 round(digits = 2)
```

The first step of a sequence of pipes can be a value, a variable, or a function including arguments. The code below shows a series of examples of different ways of achieving the same result. The examples use the function `round`, which also allows for a second argument `digits = 2`. Note that, when using the pipe operator, only the nominally second argument is provided to the function `round` – that is `round(digits = 2)`

```
# No pipe, using variables
tmp_variable_A <- 2
tmp_variable_B <- sqrt(tmp_variable_A)
round(tmp_variable_B, digits = 2)

# No pipe, using functions only
round(sqrt(2), digits = 2)

# Pipe starting from a value
```

```
2 %>%
  sqrt() %>%
  round(digits = 2)

# Pipe starting from a variable
the_value_two <- 2
the_value_two %>%
  sqrt() %>%
  round(digits = 2)

# Pipe starting from a function
sqrt(2) %>%
 round(digits = 2)
```

A complex operation created through the use of `%>%` can be used on the right side of `<-`, to assign the outcome of the operation to a variable.

```
sqrt_of_two <- 2 %>%
  sqrt() %>%
  round(digits = 2)
```

1.6 Coding style

Study the Tidyverse Style Guide (style.tidyverse.org) and use it consistently!

1.7 Exercise 104.1

Question 104.1.1: Write a piece of code using the pipe operator that takes as input the number `1632`, calculates the logarithm to the base 10, takes the highest integer number lower than the calculated value (lower round), and verifies whether it is an integer.

Question 104.1.2: Write a piece of code using the pipe operator that takes as input the number `1632`, calculates the square root, takes the lowest integer number higher than the calculated value (higher round), and verifies whether it is an integer.

Question 104.1.3: Write a piece of code using the pipe operator that takes as input the string `"1632"`, transforms it into a number, and checks whether the result is *Not a Number*.

Question 104.1.4: Write a piece of code using the pipe operator that takes as input the string `"-16.32"`, transforms it into a number, takes the absolute value and truncates it, and finally checks whether the result is *Not Available*.

1.8 Exercise 104.2

Answer the question below, consulting the `stringr` library reference (stringr.tidyverse.org/reference) as necessary

Question 104.2.1: Write a piece of code using the pipe operator and the `stringr` library that takes as input the string `"I like programming in R"`, and transforms it all in uppercase.

Question 104.2.2: Write a piece of code using the pipe operator and the `stringr` library that takes as input the string `"I like programming in R"`, and truncates it, leaving only 10 characters.

Question 104.2.3: Write a piece of code using the pipe operator and the `stringr` library that takes as input the string `"I like programming in R"`, and truncates it, leaving only 10 characters and using no ellipsis.

Question 104.2.4: Write a piece of code using the pipe operator and the `stringr` library that takes as input the string `"I like programming in R"`, and manipulates to leave only the string `"I like R"`.

Chapter 2

R programming

Stefano De Sabbata

2.1 R Scripts

The RStudio Console is handy to interact with the R interpreter and obtain results of operations and commands. However, moving from simple instructions to an actual program or scripts to conduct data analysis, the Console is usually not sufficient anymore. In fact, the Console is not a very comfortable way of providing long and complex instructions to the interpreter and editing past instructions when you want to change something. A better option to create programs or data analysis script of any significant size is to use the RStudio integrated editor to create an *R script*.

To create an R script, select from the top menu *File* > *New File* > *R Script*. That opens the embedded RStudio editor and a new empty R script folder. Copy the two lines below into the file. The first loads the `tidyverse` library, whereas the second simply calculates the square root of two.

```
# Load the Tidyverse
library(tidyverse)

# Calculate the square root of two
2 %>% sqrt()
```

```
## [1] 1.414214
```

From the top menu, select *File* > *Save*, type in *My_first_script.R* (make sure to include the underscore and the *.R* extension) as *File name*, and click *Save*. That is your first R script, congratulations!

New lines of code can be added to the file, and the whole script can then be executed. Edit the file by adding the line of code shown below, and save it. Then click the *Source* button on the top-right of the editor to execute the file. What happens the first time? What happens if you click *Source* again?

```
# First variable in a script
a_variable <- "This is my first script"
```

Alternatively, you can click on a specific line or select one or more lines, and click *Run* to execute only the selected line(s).

Delete the two lines calculating the square root of two and defining the variable `a_variable` from the script, leaving only the line loading the Tidyverse library. In the following sections, add the code to the script to execute it, rather than using the Console.

2.2 Vectors

Vectors can be defined in R by using the function `c`, which takes as parameters the items to be stored in the vector – stored in the order in which they are provided.

```
east_midlands_cities <- c("Derby", "Leicester", "Lincoln", "Nottingham")
length(east_midlands_cities)
```

```
## [1] 4
```

Once the vector has been created and assigned to an identifier, elements within the vector can be retrieved by specifying the identifier, followed by square brackets, and the *index* (or indices as we will see further below) of the elements to be retrieved – remember that indices start from 1.

```
# Retrieve the third city
east_midlands_cities[3]
```

```
## [1] "Lincoln"
```

To retrieve any subset of a vector (i.e., not just one element), specify an integer vector containing the indices of interest (rather than a single integer value) between square brackets.

```
# Retrieve first and third city
east_midlands_cities[c(1, 3)]
```

```
## [1] "Derby"   "Lincoln"
```

The operator `:` can be used to create integer vectors, starting from the number specified before the operator to the number specified after the operator.

```
# Create a vector containing integers between 2 and 4
two_to_four <- 2:4
two_to_four
```

```
## [1] 2 3 4
```

```
# Retrieve cities between the second and the fourth
east_midlands_cities[two_to_four]
```

```
## [1] "Leicester"  "Lincoln"    "Nottingham"
```

```
# As the second element of two_to_four is 3...
two_to_four[2]
```

```
## [1] 3
```

```
# the following command will retrieve the third city
east_midlands_cities[two_to_four[2]]
```

```
## [1] "Lincoln"
```

```
# Create a vector with cities from the previous vector
selected_cities <- c(east_midlands_cities[1], east_midlands_cities[3:4])
```

The functions `seq` and `rep` can also be used to create vectors, as illustrated below.

```
seq(1, 10, by = 0.5)
```

```
##  [1]  1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5  7.0  7.5  8.0
## [16]  8.5  9.0  9.5 10.0
```

```
seq(1, 10, length.out = 6)
```

```
## [1]  1.0  2.8  4.6  6.4  8.2 10.0
```

```
rep("Ciao", 4)
```

```
## [1] "Ciao" "Ciao" "Ciao" "Ciao"
```

The logical operators `any` and `all` can be used to test conditional statements on the vector. The former returns `TRUE` if at least one element satisfies the statement, the second returns `TRUE` if all elements satisfy the condition

```
any(east_midlands_cities == "Leicester")
```

```
## [1] TRUE
```

```
my_sequence <- seq(1, 10, length.out = 7)
my_sequence
```

```
## [1]  1.0  2.5  4.0  5.5  7.0  8.5 10.0
```

```
any(my_sequence > 5)
```

```
## [1] TRUE
```

```
all(my_sequence > 5)
```

```
## [1] FALSE
```

All built-in numerical functions in R can be used on a vector variable directly. That is, if a vector is specified as input, the selected function is applied to each element of the vector.

```
one_to_ten <- 1:10
one_to_ten
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
one_to_ten + 1
```

```
##  [1]  2  3  4  5  6  7  8  9 10 11
```

```
sqrt(one_to_ten)
```

```
##  [1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751 2.828427
##  [9] 3.000000 3.162278
```

2.3 Filtering

As seen in the first practical session, a conditional statement entered in the Console is evaluated for the provided input, and a logical value (`TRUE` or `FALSE`) is provided as output. Similarly, if the provided input is a vector, the conditional statement is evaluated for each element of the vector, and a vector of logical values is returned – which contains the respective results of the conditional statements for each element.

```
minus_three <- -3
minus_three > 0
```

```
## [1] FALSE
```

```
minus_three_to_three <- -3:3
minus_three_to_three
```

```
## [1] -3 -2 -1  0  1  2  3
```

```
minus_three_to_three > 0
```

```
## [1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
```

A subset of the elements of a vector can also be selected by providing a vector of logical values between brackets after the identifier. A new vector returned,

containing only the values for which a `TRUE` value has been specified correspondingly.

```
minus_two_to_two <- -2:2
minus_two_to_two
```

```
## [1] -2 -1  0  1  2
```

```
minus_two_to_two[c(TRUE, TRUE, FALSE, FALSE, TRUE)]
```

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

As the result of evaluating the conditional statement on a vector is a vector of logical values, this can be used to filter vectors based on conditional statements. If a conditional statement is provided between square brackets (after the vector identifier, instead of an index), a new vector is returned, which contains only the elements for which the conditional statement is true.

```
minus_two_to_two > 0
```

```
## [1] FALSE FALSE FALSE  TRUE  TRUE
```

```
minus_two_to_two[minus_two_to_two > 0]
```

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

2.4 Conditional statements

Conditional statements are fundamental in (procedural) programming, as they allow to execute or not execute part of a procedure depending on whether a certain condition is true. The condition is tested and the part of the procedure to execute in the case the condition is true is included in a *code block*.

```
temperature <- 25

if (temperature > 25) {
  cat("It really warm today!")
}
```

A simple conditional statement can be created using `if` as in the example above. A more complex structure can be created using both `if` and `else`, to provide not only a procedure to execute in case the condition is true, but also an alternative procedure, to be executed when the condition is false.

```
temperature <- 12

if (temperature > 25) {
  cat("It really warm today!")
} else {
```

```
  cat("Today is not warm")
}
```

```
## Today is not warm
```

Finally, conditional statements can be **nested**. That is, a conditional statement can be included as part of the code block to be executed after the condition is tested. For instance, in the example below, a second conditional statement is included in the code block to be executed in the case the condition is false.

```
temperature <- -5

if (temperature > 25) {
  cat("It really warm today!")
} else {
  if (temperature > 0) {
    cat("There is a nice temperature today")
  } else {
    cat("This is really cold!")
  }
}
```

```
## This is really cold!
```

Similarly, the first example seen in the lecture should be coded as follows.

```
a_value <- -7

if (a_value == 0) {
  cat("Zero")
} else {
  if (a_value < 0) {
    cat("Negative")
  } else {
    cat("Positive")
  }
}
```

```
## Negative
```

2.5 Loops

Loops are another core component of (procedural) programming and implement the idea of solving a problem or executing a task by performing the same set of steps a number of times. There are two main kinds of loops in R - **deterministic** and **conditional** loops. The former is executed a fixed number of times, specified at the beginning of the loop. The latter is executed until a specific condition is met. Both deterministic and conditional loops are extremely

important in working with vectors.

2.5.1 Conditional Loops

In R, conditional loops can be implemented using `while` and `repeat`. The difference between the two is mostly syntactical: the first tests the condition first and then execute the related code block if the condition is true; the second executes the code block until a `break` command is given (usually through a conditional statement).

```
a_value <- 0
# Keep printing as long as x is smaller than 2
while (a_value < 2) {
  cat(a_value, "\n")
  a_value <- a_value + 1
}
```

```
## 0
## 1
```

```
a_value <- 0
# Keep printing, if x is greater or equal than 2 than stop
repeat {
  cat(a_value, "\n")
  a_value <- a_value + 1
  if (a_value >= 2) break
}
```

```
## 0
## 1
```

2.5.2 Deterministic Loops

The deterministic loop executes the subsequent code block iterating through the elements of a provided vector. During each iteration (i.e., execution of the code block), the current element of the vector (in the definition below) is assigned to the variable in the statement (in the definition below), and it can be used in the code block.

```
for (<VAR> in <VECTOR>) {
    ... code in loop ...
    }
```

It is, for instance, possible to iterate over a vector and print each of its elements.

```
east_midlands_cities <- c("Derby", "Leicester", "Lincoln", "Nottingham")
for (city in east_midlands_cities){
  cat(city, "\n")
}
```

```
## Derby
## Leicester
## Lincoln
## Nottingham
```

It is common practice to create a vector of integers on the spot (e.g., using the : operator) to execute a certain sequence of steps a pre-defined number of times.

```
for (iterator in 1:3) {
  cat("Exectuion number", iterator, ":\n")
  cat("    Step1: Hi!\n")
  cat("    Step2: How is it going?\n")
}
```

```
## Exectuion number 1 :
##     Step1: Hi!
##     Step2: How is it going?
## Exectuion number 2 :
##     Step1: Hi!
##     Step2: How is it going?
## Exectuion number 3 :
##     Step1: Hi!
##     Step2: How is it going?
```

2.6 Exercise 114.1

Question 114.1.1: Use the modulo operator `%%` to create a conditional statement that prints `"Even"` if a number is even and `"Odd"` if a number is odd.

Question 114.1.2: Encapsulate the conditional statement written for *Question 114.1.1* into a `for` loop that executes the conditional statement for all numbers from 1 to 10.

Question 114.1.3: Encapsulate the conditional statement written for *Question 114.1.1* into a `for` loop that prints the name of cities in odd positions (i.e., first, third, fifth) in the vector `c("Birmingham", "Derby", "Leicester", "Lincoln", "Nottingham", "Wolverhampton")`.

Question 114.1.4: Write the code necessary to print the name of the cities in the vector `c("Birmingham", "Derby", "Leicester", "Lincoln", "Nottingham", "Wolverhampton")` as many times as their position in the vector (i.e., once for the first city, two times for the second, and so on and so forth).

2.7 Function definition

Recall from the lecture that *an* **algorithm** *or effective procedure is a mechanical rule, or automatic method, or programme for performing some mathematical operation* (Cutland, 1980). A **program** is a specific set of instructions that implement an abstract algorithm. The definition of an algorithm (and thus a program) can consist of one or more **function**s, which are sets of instructions that perform a task, possibly using an input, possibly returning an output value.

The code below is a simple function with one parameter. The function simply calculates the square root of a number. Add the code below to your script and run that portion of the script (or type the code into the Console).

```
cube_root <- function (input_value) {
   result <- input_value ^ (1 / 3)
   result
}
```

Once the definition of a function has been executed, the function becomes part of the environment, and it should be visible in the Environment panel, in a subsection titled *Functions*. Thereafter, the function can be called from the Console, from other portions of the script, as well as from other scripts.

If you type the instruction below in the *Console*, or add it to the script and run it, the function is called using `27` as an argument, thus returning `3`.

```
cube_root(27)
```

```
## [1] 3
```

2.7.1 Functions and control structures

One issue when writing functions is making sure that the data that has been given to the data is the right kind. For example, what happens when you try to compute the cube root of a negative number?

```
cube_root(-343)
```

```
## [1] NaN
```

That probably wasn't the answer you wanted. As you might remember `NaN` (*Not a Number*) is the value return when a mathematical expression is numerically indeterminate. In this case, this is actually due to a shortcoming with the `^` operator in R, which only works for positive base values. In fact -7 is a perfectly valid cube root of -343, since (-7)x(-7)x(-7) = -343.

To work around this limitation, we can state a conditional rule:

- If `x < 0`: calculate the cube root of x 'normally'.
- Otherwise: work out the cube root of the positive number, then change it to negative.

Those kinds of situations can be dealt with in an R function by using an `if` statement, as shown below. Note how the operator `-` (i.e., the symbol minus) is here used to obtain the inverse of a number, in the same way as `-1` is the inverse of the number `1`.

```
cube_root <- function (input_value) {
    if (input_value >= 0){
        result <- input_value^(1 / 3)
    }else{
        result <- -( (-input_value)^(1/3) )
    }
    result
}

cube_root(343)
cube_root(-343)
```

However, other things can go wrong. For example, `cube_root("Leicester")` would cause an error to occur, `Error in x^(1 / 3) : non-numeric argument to binary operator`. That shouldn't be surprising because cube roots only make sense for numbers, not character variables. Thus, it might be helpful if the cube root function could spot this and print a warning explaining the problem, rather than just crashing with a fairly cryptic error message such as the one above, as it does at the moment.

The function could be re-written to making use of `is.numeric` in a second conditional statement. If the input value is not numeric, the function returns the value `NA` (*Not Available*) instead of a number. Note that here there is an `if` statement inside another `if` statement, as it is always possible to nest code blocks – and `if` within a `for` within a `while` within an `if` within … etc.

```
cube_root <- function (input_value) {
    if (is.numeric(input_value)) {
        if (input_value >= 0){
            result <- input_value^(1/3)
        }else{
            result <- -(-input_value)^(1/3)
        }
        result
    }else{
        cat("WARNING: Input variable must be numeric\n")
        NA
    }
}
```

Finally, `cat` is a printing function, that instructs R to display the provided argument (in this case, the phrase within quotes) as output in the Console. The `\n` in `cat` tells R to add a *newline* when printing out the warning.

2.8 Exercise 114.2

Question 114.2.1: Write a function that calculates the areas of a circle, taking the radius as the first parameter.

Question 114.2.2: Write a function that calculates the volume of a cylinder, taking the radius of the base as the first parameter and the height as the second parameter. The function should call the function defined above and multiply the returned value by the height to calculate the result.

Question 114.2.3: Write a function with two parameters, a vector of numbers and a vector of characters (text). The function should check that the input has the correct data type. If all the numbers in the first vector are greater than zero, return the elements of the second vector from the first to the length of the first vector.

Chapter 3

Data wrangling Pt. 1

Stefano De Sabbata

3.1 R Projects

RStudio provides an extremely useful functionality to organise all your code and data, that is **R Projects**. Those are specialised files that RStudio can use to store all the information it has on a specific project that you are working on – *Environment*, *History*, working directory, and much more, as we will see in the coming weeks.

In RStudio Server, in the *Files* tab of the bottom-left panel, click on *Home* to make sure you are in your home folder – if you are working on your own computer, create a folder for these practicals wherever most convenient. Click on *New Folder* and enter *Practicals* in the prompt dialogue, to create a folder named *Practicals*.

Select *File > New Project …* from the main menu, then from the prompt menu, *New Directory*, and then *New Project*. Insert *Practical_204* as the directory name, and select the *Practicals* folder for the field *Create project as subdirectory of*. Finally, click *Create Project*.

RStudio has now created the project, and it should have activated it. If that is the case, the *Files* tab in the bottom-right panel should be in the *Practical_204* folder, which contains only the *Practical_204.Rproj* file. The *Practical_204.Rproj* stores all the *Environment* information for the current project and all the project files (e.g., R scripts, data, output files) should be stored within the *Practical_204* folder. Moreover, the *Practical_204* is now your working directory, which means that you can refer to a file in the folder by using

only its name and if you save a file that is the default directory where to save it.

On the top-right corner of RStudio, you should see a blue icon representing an R in a cube, next to the name of the project (*Practical_204*). That also indicates that you are within the *Practical_204* project. Click on *Practical_204* and select *Close Project* to close the project. Next to the R in a cube icon, you should now see *Project: (None)*. Click on *Project: (None)* and select *Practical_204* from the list to reactivate the *Practical_204* project.

With the *Practical_204* project activated, select from the top menu *File > New File > R Script*. That opens the embedded RStudio editor and a new empty R script folder. Copy the two lines below into the file. The first loads the `tidyverse` library, whereas the second loads another library that the code below uses to produce well-formatted tables.

```
library(tidyverse)
library(knitr)
```

From the top menu, select *File > Save*, type in *My_script_Practical_204.R* (make sure to include the underscore and the *.R* extension) as *File name*, and click *Save*.

3.2 Install libraries

RStudio and RStudio Server come with a number of libraries already pre-installed. However, you might find yourself in the position of wanting to install additional libraries to work with.

The remainder of this practical requires the library `nycflights13`. To install it, select *Tools > Install Packages...* from the top menu. Insert `nycflights13` in the *Packages (separate multiple with space or comma)* field and click install. RStudio will automatically execute the command `install.packages("nycflights13")` (so, no need to execute that yourself) and install the required library.

As usual, use the function `library` to load the newly installed library.

```
library(nycflights13)
```

The library `nycflights13` contains a dataset storing data about all the flights departed from New York City in 2013. The code below, loads the data frame `flights` from the library `nycflights13` into the variable `flights_from_nyc`, using the `::` operator to indicate that the data frame `flights` is situated within the library `nycflights13`.

```
flights_from_nyc <- nycflights13::flights
```

Add both lines above to your R script, as well as the code snippets provided as an example below.

3.2.1 Loading R scripts

It is furthermore possible to load the function(s) defined in one script from another script – in a fashion similar to when a library is loaded. Create a new R script named `Practical_204_RS_functions.R`, copy the code below in that R script and save the file

```
cube_root <- function (input_value) {
   result <- input_value ^ (1 / 3)
   result
}
```

Create a second R script named `Practical_204_RS_main.R`, copy the code below in that second R script and save the file.

```
source("Practical_204_RS_functions.R")

cube_root(27)
```

Executing the `Practical_204_RS_main.R` instructs the interpreter first to run the `Practical_204_RS_functions.R` script, thus creating the `cube_root` function, and then invoke the function using `27` as an argument, thus returning again `3`. That is a simple example, but this can be an extremely powerful tool to create your own library of functions to be used by different scripts.

3.3 Data manipulation

The analysis below uses the `dplyr` library (also part of the Tidyverse), which it offers a grammar for data manipulation.

For instance, the function `count` can be used to count the number rows of a data frame. The code below provides `flights_from_nyc` as input to the function `count` through the pipe operator, thus creating a new `tibble` with only one row and one column.

As discussed in the previous lecture, a `tibble` is data type similar to data frames, used by all the Tidyverse libraries.

All Tidyverse functions output `tibble` rather than `data.frame` objects when representing a table. However, `data.frame` object can be provided as input, as they are automatically converted by Tidyverse functions before proceeding with the processing steps.

In the `tibble` outputted by the `count` function below, the column `n` provides the count. The function `kable` of the library `knitr` is used to produce a well-formatted table.

```
flights_from_nyc %>%
  dplyr::count() %>%
  knitr::kable()
```

n
336776

The example above already shows how the **pipe operator** can be used effectively in a multi-step operation.

The function `count` can also be used to count the number rows of a table that have the same value for a given column, usually representing a category.

In the example below, the column name `origin` is provided as an argument to the function `count`, so rows representing flights from the same origin are counted together – EWR is the Newark Liberty International Airport, JFK is the John F. Kennedy International Airport, and LGA is LaGuardia Airport.

```
flights_from_nyc %>%
  dplyr::count(origin) %>%
  knitr::kable()
```

origin	n
EWR	120835
JFK	111279
LGA	104662

As you can see, the code above is formatted in a way similar to a code block, although it is not a code block. The code goes to a new line after every `%>%`, and space is added at the beginning of new lines. That is very common in R programming (especially when functions have many parameters) as it makes the code more readable.

3.3.1 Summarise

To carry out more complex aggregations, the function `summarise` can be used in combination with the function `group_by` to summarise the values of the rows of a data frame. Rows having the same value for a selected column (in the example below, the same origin) are grouped together, then values are aggregated based on the defined function (using one or more columns in the calculation).

In the example below, the function `sum` is applied to the column `distance` to calculate `distance_traveled_from` (the total distance travelled by flights starting from each airport).

```
flights_from_nyc %>%
  dplyr::group_by(origin) %>%
  dplyr::summarise(
    distance_traveled_from = sum(distance)
  ) %>%
  knitr::kable()
```

origin	distance_traveled_from
EWR	127691515
JFK	140906931
LGA	81619161

3.3.2 Select and filter

The function `select` can be used to select some **columns** to output. For instance in the code below, the function `select` is used to select the columns `origin`, `dest`, and `dep_delay`, in combination with the function `slice_head`, which can be used to include only the first `n` rows (`5` in the example below) to output.

```
flights_from_nyc %>%
  dplyr::select(origin, dest, dep_delay) %>%
  dplyr::slice_head(n = 5) %>%
  knitr::kable()
```

origin	dest	dep_delay
EWR	IAH	2
LGA	IAH	4
JFK	MIA	2
JFK	BQN	-1
LGA	ATL	-6

The function `filter` can instead be used to filter **rows** based on a specified condition. In the example below, the output of the `filter` step only includes the rows where the value of `month` is `11` (i.e., the eleventh month, November).

```
flights_from_nyc %>%
  dplyr::select(origin, dest, year, month, day, dep_delay) %>%
  dplyr::filter(month == 11) %>%
  dplyr::slice_head(n = 5) %>%
  knitr::kable()
```

origin	dest	year	month	day	dep_delay
JFK	PSE	2013	11	1	6
JFK	SYR	2013	11	1	105
EWR	CLT	2013	11	1	-5
LGA	IAH	2013	11	1	-6
JFK	MIA	2013	11	1	-3

Notice how `filter` is used in combination with `select`. All functions in the `dplyr` library can be combined, in any other order that makes logical sense. However, if the `select` step didn't include `month`, that same column couldn't have been used in the `filter` step.

3.3.3 Mutate

The function `mutate` can be used to add a new column to an output table. The `mutate` step in the code below adds a new column `air_time_hours` to the table obtained through the pipe, that is the flight air time in hours, dividing the flight air time in minutes by `60`.

```
flights_from_nyc %>%
  dplyr::select(flight, origin, dest, air_time) %>%
  dplyr::mutate(
    air_time_hours = air_time / 60
  ) %>%
  dplyr::slice_head(n = 5) %>%
  knitr::kable()
```

flight	origin	dest	air_time	air_time_hours
1545	EWR	IAH	227	3.783333
1714	LGA	IAH	227	3.783333
1141	JFK	MIA	160	2.666667
725	JFK	BQN	183	3.050000
461	LGA	ATL	116	1.933333

3.3.4 Arrange

The function `arrange` can be used to sort a tibble by ascending order of the values in the specified column. If the operator `-` is specified before the column name, the descending order is used. The code below would produce a table showing all the rows when ordered by descending order of air time.

```
flights_from_nyc %>%
  dplyr::select(flight, origin, dest, air_time) %>%
  dplyr::arrange(-air_time) %>%
  knitr::kable()
```

In the examples above, we have used `slice_head` to present only the first `n` (in the examples `5`) rows in a table, based on the existing order. The `dplyr` library also provides the functions `slice_max` and `slice_min` which incorporate the sorting functionality (see `slice` reference page).

As such, the following code uses `slice_max` to produce a table including only the 5 rows with the *highest* air time.

```
flights_from_nyc %>%
  dplyr::select(flight, origin, dest, air_time) %>%
  dplyr::slice_max(air_time, n = 5) %>%
  knitr::kable()
```

flight	origin	dest	air_time
15	EWR	HNL	695
51	JFK	HNL	691
51	JFK	HNL	686
51	JFK	HNL	686
51	JFK	HNL	683

The following code, instead, uses `slice_min`, thus producing a table including only the 5 rows with the *lowest* air time.

```
flights_from_nyc %>%
  dplyr::select(flight, origin, dest, air_time) %>%
  dplyr::slice_min(air_time, n = 5) %>%
  knitr::kable()
```

flight	origin	dest	air_time
4368	EWR	BDL	20
4631	EWR	BDL	20
4276	EWR	BDL	21
4619	EWR	PHL	21
4368	EWR	BDL	21
4619	EWR	PHL	21
2132	LGA	BOS	21
3650	JFK	PHL	21
4118	EWR	BDL	21
4276	EWR	BDL	21
4276	EWR	BDL	21
4276	EWR	BDL	21
4276	EWR	BDL	21
4577	EWR	BDL	21
6062	EWR	BDL	21
3847	EWR	BDL	21

In both cases, if the table contains ties, all rows containing a value that is present among the maximum or minimum selected values are presented, as it is the case with the rows containing the value `21` in the example above.

3.4 Data manipulation example

Finally, the code below illustrates a more complex, multi-step operation using all the functions discussed above.

1. Start from the `flights_from_nyc` data.
2. Select origin, destination, departure delay, year, month, and day.
3. Filter only rows referring to flights in November.
4. Filter only rows where departure delay is not (notice that the negation operator `!` is used) `NA`.

- That is necessary because the function `mean` would return `NA` as output if any of the values in the column is `NA`.

5. Group by destination.
6. Calculated the average delay per destination.
7. Add a column with the delay calculated in hours (minutes over 60).
8. Sort the table by *descending* delay (note that `-` is used before the column name).
9. Only show the first 5 rows.
10. Create a well-formatted table.

```
flights_from_nyc %>%
  dplyr::select(origin, dest, year, month, day, dep_delay) %>%
  dplyr::filter(month == 11) %>%
  dplyr::filter(!is.na(dep_delay)) %>%
  dplyr::group_by(dest) %>%
  dplyr::summarize(
    avg_dep_delay = mean(dep_delay)
  ) %>%
  dplyr::mutate(
    avg_dep_delay_hours = avg_dep_delay / 60
  ) %>%
  dplyr::arrange(-avg_dep_delay_hours) %>%
  dplyr::slice_head(n = 5) %>%
  knitr::kable()
```

dest	avg_dep_delay	avg_dep_delay_hours
SBN	67.50000	1.1250000
BDL	26.66667	0.4444444
CAK	19.70909	0.3284848
BHM	19.61905	0.3269841
DSM	16.14815	0.2691358

3.5 Exercise 204.1

Extend the code in the script `My_script_Practical_204.R` to include the code necessary to solve the questions below.

Question 204.1.1: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the average air time in hours, calculated grouping flights by carrier, but only for flights starting from the JFK airport.

Question 204.1.2: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the average arrival delay compared to the overall air time (**tip**: use manipulate to create a new column that takes the result of `arr_delay / air_time`) calculated grouping flights by carrier, but only for flights starting from the JFK airport.

Question 204.1.3: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the average arrival delay compared to the overall air time calculated grouping flights by origin and destination, sorted by destination.

Chapter 4

Data wrangling Pt. 2

Stefano De Sabbata

This section illustrates the re-shape and join functionalities of the Tidyverse libraries using simple examples. The following sections instead present a more complex example, loading and wrangling with data related to the 2011 Output Area Classification and the Indexes of Multiple Deprivation 2015.

```
library(tidyverse)
library(knitr)
```

4.1 Table manipulation

4.1.1 Long and wide formats

Tabular data are usually presented in two different formats.

- **Wide**: this is the most common approach, where each real-world entity (e.g. a city) is represented by *one single row* and its attributes are represented through different columns (e.g., a column representing the total population in the area, another column representing the size of the area, etc.).

City	Population	Area	Density
Leicester	329,839	73.3	4,500
Nottingham	321,500	74.6	4,412

- **Long**: this is probably a less common approach, but still necessary in

many cases, where each real-world entity (e.g. a city) is represented by *multiple rows*, each one reporting only one of its attributes. In this case, one column is used to indicate which attribute each row represent, and another column is used to report the value.

City	Attribute	Value
Leicester	Population	329,839
Leicester	Area	73.3
Leicester	Density	4,500
Nottingham	Population	321,500
Nottingham	Area	74.6
Nottingham	Density	4,412

The `tidyr` library provides two functions that allow transforming wide-formatted data to a long format, and vice-versa. Please take your time to understand the example below and check out the `tidyr` help pages before continuing.

```
city_info_wide <- data.frame(
    city = c("Leicester", "Nottingham"),
    population = c(329839, 321500),
    area = c(73.3, 74.6),
    density = c(4500, 4412)
  ) %>%
  tibble::as_tibble()

city_info_wide %>%
 knitr::kable()
```

city	population	area	density
Leicester	329839	73.3	4500
Nottingham	321500	74.6	4412

```
city_info_long <- city_info_wide %>%
  tidyr::pivot_longer(
    # exclude IDs (city names) from the pivoted columns
    cols = -city,
    # name for the new column containing
    # the names of the old columns
    names_to = "attribute",
    # name for the new column containing
    # the values included under the old columns
    values_to = "value"
  )

city_info_long %>%
```

```
  knitr::kable()
```

city	attribute	value
Leicester	population	329839.0
Leicester	area	73.3
Leicester	density	4500.0
Nottingham	population	321500.0
Nottingham	area	74.6
Nottingham	density	4412.0

```
city_info_back_to_wide <- city_info_long %>%
  tidyr::pivot_wider(
    # column containing the attribute names
    names_from = attribute,
    # column containing the values
    values_from = value
  )

city_info_back_to_wide %>%
  knitr::kable()
```

city	population	area	density
Leicester	329839	73.3	4500
Nottingham	321500	74.6	4412

4.1.2 Join

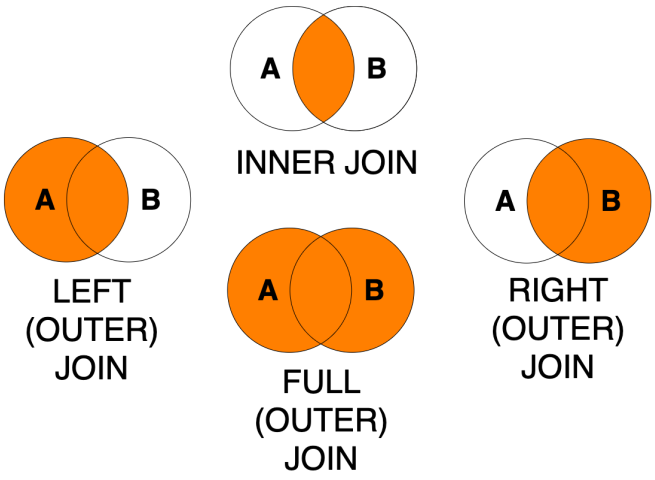
A join operation combines two tables into one by matching rows that have the same values in the specified column. This operation is usually executed on columns containing identifiers, which are matched through different tables containing different data about the same real-world entities. For instance, the table below presents the telephone prefixes for two cities. That information can be combined with the data present in the wide-formatted table above through a join operation on the columns containing the city names. As the two tables do not contain all the same cities, if a full join operation is executed, some cells have no values assigned.

city	telephone_prefix
Leicester	0116
Birmingham	0121

city	population	area	density	telephone_prefix
Leicester	329,839	73.3	4,500	0116
Nottingham	321,500	74.6	4,412	

city	population	area	density	telephone_prefix
Birmingham				0121

As discussed in the lecture, the `dplyr` library offers different types of join operations, which correspond to the different `SQL` joins illustrated in the image below. The use and implications of these different types of joins will be discussed in more detail in the GY7708 module next semester.



Please take your time to understand the example below and check out the related `dplyr` help pages before continuing. The first four examples execute the exact same *full join* operation using three different syntaxes: with or without using the pipe operator, and specifying the `by` argument or not. Note that all those approaches to writing the join are valid and produce the same result. The choice about which approach to use will depend on the code you are writing. In particular, you might find it useful to use the syntax that uses the pipe operator when the join operation is itself only one stem in a series of data manipulation steps. Using the `by` argument is usually advisable unless you are certain that you aim to join two tables with all and exactly the column that have the same names in the two table.

Note how the result of the join operations is *not* saved to a variable. The function `knitr::kable` is added after each join operation through a pipe `%>%` to display the resulting table in a nice format.

```
city_telephone_prexix <- data.frame(
    city = c("Leicester", "Birmingham"),
    telephon_prefix = c("0116", "0121")
  ) %>%
  tibble::as_tibble()
```

```
city_telephone_prexix %>%
  knitr::kable()
```

city	telephon_prefix
Leicester	0116
Birmingham	0121

```
# Option 1: without using the pipe operator

# full join verb
dplyr::full_join(
    # left table
    city_info_wide,
    # right table
    city_telephone_prexix,
    # columns to match
    by = c("city" = "city")
  ) %>%
  knitr::kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Nottingham	321500	74.6	4412	NA
Birmingham	NA	NA	NA	0121

```
# Option 2: without using the pipe operator
#   and without using the argument "by"
#   as columns have the same name
#   in the two tables.
# Same result as Option 1

# full join verb
dplyr::full_join(
    # left table
    city_info_wide,
    # right table
    city_telephone_prexix
  ) %>%
  knitr::kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Nottingham	321500	74.6	4412	NA
Birmingham	NA	NA	NA	0121

```
# Option 3: using the pipe operator
#   and without using the argument "by"
#   as columns have the same name
#   in the two tables.
# Same result as Option 1 and 2

# left table
city_info_wide %>%
  # full join verb
  dplyr::full_join(
    # right table
    city_telephone_prexix
  ) %>%
  knitr::kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Nottingham	321500	74.6	4412	NA
Birmingham	NA	NA	NA	0121

```
# Option 4: using the pipe operator
#   and using the argument "by".
# Same result as Option 1, 2 and 3

# left table
city_info_wide %>%
  # full join verb
  dplyr::full_join(
    # right table
    city_telephone_prexix,
    # columns to match
    by = c("city" = "city")
  ) %>%
  knitr::kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Nottingham	321500	74.6	4412	NA
Birmingham	NA	NA	NA	0121

```
# Left join
# Using syntax similar to Option 1 above

# left join
dplyr::left_join(
    # left table
    city_info_wide,
    # right table
    city_telephone_prexix,
    # columns to match
    by = c("city" = "city")
  ) %>%
  kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Nottingham	321500	74.6	4412	NA

```
# Right join
# Using syntax similar to Option 2 above

# right join verb
dplyr::right_join(
    # left table
    city_info_wide,
    # right table
    city_telephone_prexix
  ) %>%
  kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116
Birmingham	NA	NA	NA	0121

```r
# Inner join
# Using syntax similar to Option 3 above

# left table
city_info_wide %>%
  # inner join
  dplyr::inner_join(
    # right table
    city_telephone_prexix
  ) %>%
  kable()
```

city	population	area	density	telephon_prefix
Leicester	329839	73.3	4500	0116

4.2 Read and write data

The `readr` library (also part of the Tidyverse) provides a series of functions that can be used to load from and save data to different file formats.

Download from Blackboard (or the data folder of the repository) the following files:

- `2011_OAC_Raw_uVariables_Leicester.csv`
- `IndexesMultipleDeprivation2015_Leicester.csv`

Create a *Practical_214* project and make sure it is activated and thus the *Practical_214* showing in the *File* tab in the bottom-right panel. Upload the two files to the *Practical_214* folder by clicking on the *Upload* button and selecting the files from your computer (at the time of writing, Chrome seems to upload the file correctly, whereas it might be necessary to change the names of the files after upload using Microsoft Edge).

Create a new R script named `Data_Wrangling_Example.R` in the *Practical_214* project, and add `library(tidyverse)` as the first line. Use that new script for this and the following sections of this practical session.

The 2011 Output Area Classification (2011 OAC) is a geodemographic classification of the census Output Areas (OA) of the UK, which was created by Gale et al. (2016) starting from an initial set of 167 prospective variables from the United Kingdom Census 2011: 86 were removed, 41 were retained as they are, and 40 were combined, leading to a final set of 60 variables. Gale et al. (2016) finally used the k-means clustering approach to create 8 clusters or supergroups (see map at datashine.org.uk), as well as 26 groups and 76 subgroups. The dataset in the file `2011_OAC_Raw_uVariables_Leicester.csv` contains all the original 167 variables, as well as the resulting groups, for the city of Leicester. The full variable names can be found in the file `2011_OAC_Raw_uVariables_Lookup.csv`.

The Indexes of Multiple Deprivation 2015 (see map at cdrc.ac.uk) are based on a series of variables across seven distinct domains of deprivation which are combined to calculate the Index of Multiple Deprivation 2015 (IMD 2015). That is an overall measure of multiple deprivations experienced by people living in an area. These indexes are calculated for every Lower layer Super Output Area (LSOA), which are larger geographic unit than the OAs used for the 2011 OAC. The dataset in the file `IndexesMultipleDeprivation2015_Leicester.csv` contains the main Index of Multiple Deprivation, as well as the values for the seven distinct domains of deprivation, and two additional indexes regarding deprivation affecting children and older people. The dataset includes scores, ranks (where 1 indicates the most deprived area), and decile (i.e., the first decile includes the 10% most deprived areas in England).

The `read_csv` function reads a *Comma Separated Values (CSV)* file from the path provided as the first argument. The code below loads the 2011 OAC dataset. The `read_csv` instruction throws a warning that shows the assumptions about the data types used when loading the data. As illustrated by the output of the last line of code, the data are loaded as a tibble 969 x 190, that is 969 rows – one for each OA – and 190 columns, 167 of which represent the input variables used to create the 2011 OAC.

```
leicester_2011OAC <-
  readr::read_csv("2011_OAC_Raw_uVariables_Leicester.csv")

leicester_2011OAC %>%
  dplyr::select(OA11CD,LSOA11CD, supgrpcode,supgrpname,Total_Population) %>%
  dplyr::slice_head(n = 3) %>%
  knitr::kable()
```

OA11CD	LSOA11CD	supgrpcode	supgrpname	Total_Population
E00069517	E01013785	6	Suburbanites	313
E00069514	E01013784	2	Cosmopolitans	323
E00169516	E01013713	4	Multicultural Metropolitans	341

The code below loads the IMD 2015 dataset.

```
# Load Indexes of Multiple deprivation data
leicester_IMD2015 <-
  readr::read_csv("IndexesMultipleDeprivation2015_Leicester.csv")
```

The function `write_csv` can be used to save a dataset as a `csv` file. For instance, the code below uses `tidyverse` functions and the pipe operator `%>%` to:

1. **read** the 2011 OAC dataset again directly from the file, but without storing it into a variable;
2. **select** the OA code variable `OA11CD`, and the two variables representing the code and name of the supergroup assigned to each OA by the 2011 OAC (`supgrpcode` and `supgrpname` respectively);
3. **filter** only those OA in the supergroup *Suburbanites* (code `6`);

4. **write** the results to a file named *Leicester_Suburbanites.csv*.

```
readr::read_csv("2011_OAC_Raw_uVariables_Leicester.csv") %>%
  dplyr::select(OA11CD, supgrpcode, supgrpname) %>%
  dplyr::filter(supgrpcode == 6) %>%
  readr::write_csv("Leicester_Suburbanites.csv")
```

4.2.1 File paths

File paths can be specified in two different ways:

- **Absolute file path**: the full file path, from the *root* folder of your computer to the file.
  - The absolute file path of a file can be obtained using the `file.choose()` instruction from the *R Console*, which will open an interactive window that will allow you to select a file from your computer. The absolute path to that file will be printed to console.
  - Absolute file paths provide a direct link to a specific file and ensure that you are loading that exact file.
  - However, absolute file paths can be problematic if the file is moved, or if the script is run on a different system, and the file path would then be invalid
- **Relative file path**: a partial path, from the current working folder to the file.
  - The current *working directory* (current folder) is part of the environment of the `R` session and can be identified using the `getwd()` instruction from the *`*R Console*. - When a new R session is started, the current *working directory* is usually the computer user's home folder. - When working within an R project, the current *working directory* is the project directory. - The current working can be manually set to a specific directory using the function`setwd‘.
  - Using a relative path while working within an R project is the option that provides the best overall **consistency**, assuming that all (data) files to be read by scripts of a project are also contained in the project folder (or subfolder).

```
# Absolute file path
# Note: the fist / indicates the root folder
readr::read_csv("/home/username/GY7702/data/2011_OAC_Raw_uVariables_Leicester.csv")

# Relative file path
# assuming the working directory is the user home folder
# /home/username
# Note: no initial / for relative file paths
readr::read_csv("GY7702/data/2011_OAC_Raw_uVariables_Leicester.csv")
```

```
# Relative file path
# assuming you are working within and R project created in the folder
# /home/username/GY7702
# Note: no initial / for relative file paths
readr::read_csv("data/2011_OAC_Raw_uVariables_Leicester.csv")
```

4.3 Data wrangling example

4.3.1 Re-shaping

The IMD 2015 data are in a *long* format, which means that every area is represented by more than one row: the column `Value` presents the value; the column `IndicesOfDeprivation` indicates which index the value refers to; the column `Measurement` indicates whether the value is a score, rank, or decile. The code below illustrates the data format used for the `IndicesOfDeprivation` table, and showing the rows for the LSOA including the University of Leicester (feature code `E01013649`).

```
leicester_IMD2015 %>%
  dplyr::filter(FeatureCode == "E01013649") %>%
  dplyr::select(FeatureCode, IndicesOfDeprivation, Measurement, Value) %>%
  knitr::kable()
```

FeatureCode	IndicesOfDeprivation	Measurement	Value
E01013649	Income Deprivation Domain	Score	0.070
E01013649	Employment Deprivation Domain	Score	0.075
E01013649	Income Deprivation Affecting Children Index (IDACI)	Score	0.087
E01013649	Income Deprivation Affecting Older People Index (IDAOPI)	Score	0.153
E01013649	Health Deprivation and Disability Domain	Score	0.272
E01013649	Index of Multiple Deprivation (IMD)	Score	19.665
E01013649	Education, Skills and Training Domain	Score	2.195
E01013649	Barriers to Housing and Services Domain	Score	14.324
E01013649	Living Environment Deprivation Domain	Score	57.197
E01013649	Crime Domain	Score	1.159
E01013649	Income Deprivation Domain	Rank	23511.000
E01013649	Index of Multiple Deprivation (IMD)	Rank	14539.000
E01013649	Employment Deprivation Domain	Rank	21227.000
E01013649	Education, Skills and Training Domain	Rank	30744.000
E01013649	Barriers to Housing and Services Domain	Rank	23885.000
E01013649	Health Deprivation and Disability Domain	Rank	12269.000
E01013649	Living Environment Deprivation Domain	Rank	1197.000
E01013649	Crime Domain	Rank	2214.000
E01013649	Income Deprivation Affecting Children Index (IDACI)	Rank	22984.000
E01013649	Income Deprivation Affecting Older People Index (IDAOPI)	Rank	16055.000
E01013649	Employment Deprivation Domain	Decile	7.000
E01013649	Income Deprivation Affecting Older People Index (IDAOPI)	Decile	5.000
E01013649	Barriers to Housing and Services Domain	Decile	8.000
E01013649	Income Deprivation Affecting Children Index (IDACI)	Decile	7.000
E01013649	Crime Domain	Decile	1.000
E01013649	Income Deprivation Domain	Decile	8.000
E01013649	Health Deprivation and Disability Domain	Decile	4.000
E01013649	Living Environment Deprivation Domain	Decile	1.000
E01013649	Education, Skills and Training Domain	Decile	10.000
E01013649	Index of Multiple Deprivation (IMD)	Decile	5.000

In the following section, the analysis aims to explore how certain census variables vary in areas with different deprivation levels. Thus, we need to extract the `Decile` rows from the IMD 2015 dataset and transform the data in a *wide* format, where each index is represented as a separate column.

To that purpose, we also need to change the name of the indexes slightly, to exclude spaces and punctuation, so that the new column names are simpler than the original text, and can be used as column names. That part of the manipulation is performed using `mutate` and functions from the `stringr` library.

```
leicester_IMD2015_decile_wide <- leicester_IMD2015 %>%
  # Select only Socres
  dplyr::filter(Measurement == "Decile") %>%
  # Trim names of IndicesOfDeprivation
  dplyr::mutate(
    IndicesOfDeprivation = str_replace_all(IndicesOfDeprivation, "\\s", "")
  ) %>%
  dplyr::mutate(
    IndicesOfDeprivation = str_replace_all(IndicesOfDeprivation, "[:punct:]", "")
  ) %>%
  dplyr::mutate(
    IndicesOfDeprivation = str_replace_all(IndicesOfDeprivation, "\\(", "")
  ) %>%
  dplyr::mutate(
    IndicesOfDeprivation = str_replace_all(IndicesOfDeprivation, "\\)", "")
  ) %>%
  # Spread
  pivot_wider(
    names_from = IndicesOfDeprivation,
    values_from = Value
  ) %>%
  # Drop columns
  dplyr::select(-DateCode, -Measurement, -Units)
```

Let's compare the columns of the original *long* IMD 2015 dataset with the *wide* dataset created above, using the function `colnames`.

```
leicester_IMD2015 %>%
  colnames() %>%
  # limit width of printing area
  print(width = 70)
```

```
## [1] "FeatureCode"          "DateCode"
## [3] "Measurement"          "Units"
## [5] "Value"                "IndicesOfDeprivation"
```

```
leicester_IMD2015_decile_wide %>%
  colnames() %>%
  # limit width of printing area
  print(width = 70)
```

```
##  [1] "FeatureCode"
##  [2] "HealthDeprivationandDisabilityDomain"
##  [3] "IncomeDeprivationAffectingOlderPeopleIndexIDAOPI"
##  [4] "BarrierstoHousingandServicesDomain"
##  [5] "EmploymentDeprivationDomain"
##  [6] "EducationSkillsandTrainingDomain"
```

```
##  [7] "LivingEnvironmentDeprivationDomain"
##  [8] "IncomeDeprivationAffectingChildrenIndexIDACI"
##  [9] "CrimeDomain"
## [10] "IndexofMultipleDeprivationIMD"
## [11] "IncomeDeprivationDomain"
```

In `leicester_IMD2015_decile_wide`, we now have only one row representing the LSOA including the University of Leicester (feature code `E01013649`) and the main Index of Multiple Deprivations is now represented by the column `IndexofMultipleDeprivationIMD`. The value reported is the same – that is 5, which means that the selected LSOA is estimated to be in the range 40-50% most deprived areas in England – but we changed the data format.

```
# Original long IMD 2015 dataset
leicester_IMD2015 %>%
  dplyr::filter(
    FeatureCode == "E01013649",
    IndicesOfDeprivation == "Index of Multiple Deprivation (IMD)",
    Measurement == "Decile"
  ) %>%
  dplyr::select(FeatureCode, IndicesOfDeprivation, Measurement, Value) %>%
  knitr::kable()
```

FeatureCode	IndicesOfDeprivation	Measurement	Value
E01013649	Index of Multiple Deprivation (IMD)	Decile	5

```
# New wide IMD 2015 dataset
leicester_IMD2015_decile_wide %>%
  dplyr::filter(FeatureCode == "E01013649") %>%
  dplyr::select(FeatureCode, IndexofMultipleDeprivationIMD) %>%
  knitr::kable()
```

FeatureCode	IndexofMultipleDeprivationIMD
E01013649	5

4.3.2 Join

As discussed above, two tables can be joined using a common column of identifiers. We can thus join the 2011 OAC and the IMD 2015 datasets into a single table. The LSOA code included in the 2011 OAC table is used to match that information with the corresponding row in the IMD 2015. The resulting table provides all the information from the 2011 OAC for each OA, plus the Index of Multiple Deprivations decile for the LSOA containing each OA.

That operation can be carried out using the function `inner_join`, and specifying the common column (or columns, if more than one is to be used as identifier) as argument of `by`. Note that using `inner_join` would result in dropping any row which doesn't have a match in the other table, either way. In this case, that should not happen, as all OAs are part of an LSOA, and any LSOA contains at

least one OA.

```
leicester_2011OAC_IMD2015 <-
  leicester_2011OAC %>%
  inner_join(
    leicester_IMD2015_decile_wide,
    by = c("LSOA11CD" = "FeatureCode")
  )
```

As each LSOA contains multiple OAs, each row from the `leicester_IMD2015_decile_wide` table is matched to multiple rows from the `leicester_2011OAC` table. For instance, as shown in the table below, the information from the IMD 2015 dataset about the LSOA encompassing the University of Leicester (feature code `E01013649`) is joined to multiple rows from the 2011 OAC dataset, including the OA encompassing the University of Leicester (feature code `E00068890`) as well as other neighbouring OAs.

```
leicester_2011OAC_IMD2015 %>%
  # Note that the LSOA11CD column needs to be used
  # as the previous join as combined
  # LSOA11CD and FeatureCode
  # into one, name LSOA11CD
  dplyr::filter(LSOA11CD == "E01013649") %>%
  dplyr::select(OA11CD, LSOA11CD, supgrpname, IndexofMultipleDeprivationIMD) %>%
  knitr::kable()
```

OA11CD	LSOA11CD	supgrpname	IndexofMultipleDeprivationIMD
E00169447	E01013649	Cosmopolitans	5
E00168083	E01013649	Cosmopolitans	5
E00068893	E01013649	Cosmopolitans	5
E00068892	E01013649	Cosmopolitans	5
E00068890	E01013649	Cosmopolitans	5

Once the result is stored into the variable `leicester_2011OAC_IMD2015`, further analysis can be carried out. For instance, `count` can be used to count how many OAs fall into each 2011 OAC supergroup and decile of the Index of Multiple Deprivations.

```
leicester_2011OAC_IMD2015 %>%
  dplyr::count(supgrpname, IndexofMultipleDeprivationIMD) %>%
  knitr::kable()
```

supgrpname	IndexofMultipleDeprivationIMD	n
Constrained City Dwellers	1	30
Constrained City Dwellers	2	3
Constrained City Dwellers	3	2
Constrained City Dwellers	6	1
Cosmopolitans	2	25
Cosmopolitans	3	15
Cosmopolitans	4	15
Cosmopolitans	5	8
Cosmopolitans	6	10
Cosmopolitans	8	10
Ethnicity Central	1	28
Ethnicity Central	2	18
Ethnicity Central	3	5
Ethnicity Central	4	5
Ethnicity Central	5	1
Hard-Pressed Living	1	68
Hard-Pressed Living	2	24
Hard-Pressed Living	3	5
Hard-Pressed Living	5	3
Hard-Pressed Living	6	1
Multicultural Metropolitans	1	107
Multicultural Metropolitans	2	119
Multicultural Metropolitans	3	132
Multicultural Metropolitans	4	100
Multicultural Metropolitans	5	56
Multicultural Metropolitans	6	34
Multicultural Metropolitans	7	6
Multicultural Metropolitans	8	17
Multicultural Metropolitans	9	2
Suburbanites	2	2
Suburbanites	3	2
Suburbanites	4	2
Suburbanites	5	10
Suburbanites	6	11
Suburbanites	7	16
Suburbanites	8	3
Suburbanites	9	8
Urbanites	1	1
Urbanites	2	3
Urbanites	3	5
Urbanites	4	15
Urbanites	5	14
Urbanites	6	4
Urbanites	7	12
Urbanites	8	7
Urbanites	9	4

As another example, the code below can be used to group OAs based on the decile and then calculate the percentage of adults not in employment using the `u074` (*No adults in employment in household: With dependent children*) and `u075` (*No adults in employment in household: No dependent children*) variables from the 2011 OAC dataset.

```
leicester_2011OAC_IMD2015 %>%
  dplyr::group_by(IndexofMultipleDeprivationIMD) %>%
  dplyr::summarise(
    adults_not_empl_perc = (sum(u074 + u075) / sum(Total_Population)) * 100
  ) %>%
  knitr::kable()
```

IndexofMultipleDeprivationIMD	adults_not_empl_perc
1	17.071876
2	14.191205
3	10.405029
4	9.966309
5	11.337036
6	10.710509
7	10.641026
8	9.686658
9	9.898140

4.4 Exercise 214.1

Extend the code in the script `Data_Wrangling_Example.R` to include the code necessary to solve the questions below. Use the full list of variable names from the 2011 UK Census used to generate the 2011 OAC thatcan be found in the file `2011_OAC_Raw_uVariables_Lookup.csv` to indetify which columns to use to complete the tasks.

Question 214.1.1: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the percentage of EU citizens over total population, calculated grouping OAs by the related decile of the Index of Multiple Deprivations, but only accounting for areas classified as Cosmopolitans or Ethnicity Central or Multicultural Metropolitans.

Question 214.1.2: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the percentage of EU citizens over total population, calculated grouping OAs by the related supergroup in the 2011 OAC, but only accounting for areas in the top 5 deciles of the Index of Multiple Deprivations.

Question 214.1.3: Write a piece of code using the pipe operator and the `dplyr` library to generate a table showing the percentage of people aged 65 and above, calculated grouping OAs by the related supergroup in the 2011 OAC and decile

of the Index of Multiple Deprivations, and ordering the table by the calculated value in a descending order.

4.5 Exercise 214.2

Extend the code in the script `Data_Wrangling_Example.R` to include the code necessary to solve the questions below.

Question 214.2.1: Write a piece of code using the pipe operator and the `dplyr` and `tidyr` libraries to generate a long format of the `leicester_2011OAC_IMD2015` table only including the values (census variables) used in *Question 214.1.3*.

Question 214.2.2: Write a piece of code using the pipe operator and the `dplyr` and `tidyr` libraries to generate a table similar to the one generated for *Question 214.2.1*, but showing the values as percentages over total population.

Chapter 5

Reproducibility

Stefano De Sabbata

5.1 Markdown

A essential tool used in creating these materials is RMarkdown That is an R library that allows you to create scripts that mix the Markdown mark-up language and R, to create dynamic documents. RMarkdown script can be compiled, at which point, the Markdown notation is interpreted to create the output files, while the R code is executed and the output incorporated in the document.

For instance the following markdown code

```
[This is a link to the University of Leicester](http://le.ac.uk) and **this is in bold**.
```

is rendered as

This is a link to the University of Leicester and **this is in bold**.

The core Markdown notation used in this session is presented below. A full RMarkdown *cheatsheet* is available here.

```
# Header 1
## Header 2
### Header 3
#### Header 4
##### Header 5

**bold**
*italics*
```

```
[This is a link to the University of Leicester](http://le.ac.uk)

- Example list
    - Main folder
        - Analysis
        - Data
        - Utils
    - Other bullet point
- And so on
    - and so forth

1. These are
    1. Numeric bullet points
    2. Number two
2. Another number two
3. This is number three
```

5.1.1 R Markdown

R code can be embedded in RMarkdown documents as in the example below. That results in the code chunk be displayed within the document (as *echo=TRUE* is specified), followed by the output from the execution of the same code.

````
```{r, echo=TRUE}
a_number <- 0
a_number <- a_number + 1
a_number <- a_number + 1
a_number <- a_number + 1
a_number
```
````

```r
a_number <- 0
a_number <- a_number + 1
a_number <- a_number + 1
a_number <- a_number + 1
a_number
```

```
## [1] 3
```

5.2 Exercise 224.1

Create a new R project named *Practical_224* as the directory name. Create an RMarkdown document in RStudio by selecting *File > New File > R Markdown …* – this might prompt RStudio to update some packages. On the RMarkdown

document creation menu, specify "Practical 05" as title and your name as the author, and select *PDF* as default output format.

The new document should contain the core document information, as in the example below, plus some additional content that simply explains how RMarkdown works.

```
---
title: "Practical 224"
author: "A. Student"
date: "7 October 2018"
output: pdf_document
---
```

Delete the contents below the document information and copy the following text below the document information.

````
# Pipe example

This is my first [RMarkdown](https://rmarkdown.rstudio.com/) document.

```{r, echo=TRUE}
library(tidyverse)
```

The code uses the pipe operator:

- takes 2 as input
- calculates the square root
- rounds the value
    - keeping only two digits

```{r, echo=TRUE}
2 %>%
  sqrt() %>%
  round(digits = 2)
```
````

The option `echo=TRUE` tells RStudio to include the code in the output document, along with the output of the computation. If `echo=FALSE` is specified, the code will be omitted. If the option `message=FALSE` and `warning=FALSE` are added, messages and warnings from R are not displayed in the output document.

Save the document by selecting *File* > *Save* from the main menu. Enter *Square_root* as file name and click *Save*. The file is saved using the *Rmd* (RMarkdown) extension.

Click on the *Knit* button on the bar above the editor panel (top-left area) in RStudio, on the left side. Check the resulting *pdf* document. Try adding some

of your own code (e.g., using some of the examples above) and Markdown text, and compile the document again.

5.3 Exercise 224.2

Create an analysis document based on RMarkdown for each one of the two analyses seen in the practical sessions 3 and 4. For each of the two analyses, within their respective R projects, first, create an RMarkdown document. Then, add the code from the related R script. Finally add additional content such as title, subtitles, and most importantly, some text describing the data used, how the analysis has been done, and the result obtained. Make sure you add appropriate links to the data sources, as available in the practical session materials.

5.4 Git

Git is a free and opensource version control system. It is commonly used through a server, where a master copy of a project is kept, but it can also be used locally. Git allows storing versions of a project, thus providing file synchronisation, consistency, history browsing, and the creation of multiple branches. For a detailed introduction to Git, please refer to the Pro Git book, written by Scott Chacon and Ben Straub.

As illustrated in the image below, when working with a git repository, the most common approach is to first check-out the latest version from the main repository before start working on any file. Once a series of edits have been made, the edits to stage are selected and then committed in a permanent snapshot. One or more commits can then be pushed to the main repository.

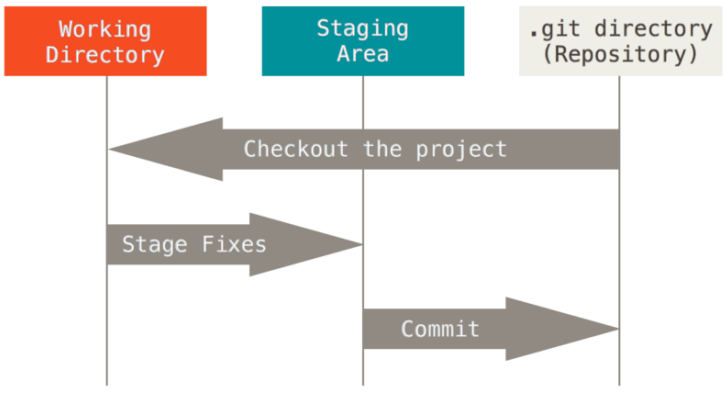


Figure 5.1: by Scott Chacon and Ben Straub, licensed under CC BY-NC-SA 3.0

5.4.1 Git and RStudio

In RStudio Server, in the *Files* tab of the bottom-left panel, click on *Home* to make sure you are in your home folder – if you are working on your own computer, create a folder for granolarr wherever most convenient. Click on *New Folder* and enter *Repos* (short for repositories) in the prompt dialogue, to create a folder named *Repos*.

Create a GitHub account at github.com, if you don't have one, and create a new repository named *my-granolarr*, following the instructions available on the GitHub help pages. Make sure you tick the box next to *Initialize this repository with a README*, which adds a `README.md` markdown file to your repository.

Once the repository has been created, GitHub will take you to the repository page. Copy the link to the repository `.git` file by clicking on the green *Clone or download* button and copying the `https` URL there available. Back to RStudio Server, select *File > New Project…* from the top menu and select *Version Control* and then *Git* from the *New Project* panel. Paste the copied URL in the *Repository URL* field, select the *Repos* folder created above as folder for *Create project as subdirectory of*, and click *Create Project*. RStudio might ask you for your GitHub username and password at this point.

Before continuing, you need to record your identity with the Git system installed on the RStudio Server, for it to be able to communicate with the GitHub's server. Open the *Terminal* tab in RStudio (if not visible, select *Tools > Terminal > New Terminal* from the top menu). First, paste the command below substituting `you@example.com` with your university email (make sure to maintain the double quotes) and press the return button.

```
git config --global user.email "you@example.com"
```

Then, paste the command below substituting `Your Name` with your name (make sure to maintain the double quotes) and press the return button.

```
git config --global user.name "Your Name"
```

RStudio should have now switched to a new R project linked to you *my-granolarr* repository. In the RStudio *File* tab in the bottom-right panel, navigate to the file created for the exercises above, select the files and copy them in the folder of the new project (in the *File* tab, *More > Copy To…*).

Check the now available *Git* tab on the top-right panel, and you should see at least the newly copied files marked as untracked. Tick the checkboxes in the *Staged* column to stage the files, then click on the *Commit* button.

In the newly opened panel *Commit* window, the top-left section shows the files, and the bottom section shows the edits. Write `My first commit` in the *Commit message* section in the top-right, and click the *Commit* button. A pop-up should notify the completed commit. Close both the pop-up panel, and click the *Push* button on the top-right of the *Commit* window. Another pop-up panel should

ask you for your GitHub username and password and then show the executed push operation. Close both the pop-up panel and the *Commit* window.

Congratulations, you have completed your first commit! Check the repository page on GitHub. If you reload the page, the top bar should show *2 commits* on the left and your files should now be visible in the file list below. If you click on *2 commits*, you can see the commit history, including both the initial commit that created the repository and the commit you just completed.

5.5 Exercise 224.3

Create a new GitHub repository named `GY7702_224_Exercise3`. Clone the repository to RStudio Server as a new an R project from that repository. Create an RMarkdown document exploring the presence of the different living arrangements in Leicester among both the different categories of the 2011 Output Area Classification and deciles of Index of Multiple Deprivations, copying the required data into the project folder.

Your analysis should include:

- an introduction to the data and the aims of the project;
- a justification of the analysis methods;
- the code and related results;
- and a discussion of the results within the same document.

5.6 Cloning granolarr

You can follow the steps listed below to clone the granolarr repository.

1. Create a folder named `Repos` in your home directory. If you are working on RStudio Server, in the *Files* panel, click on the `Home` button (second bar, next to the house icon), then click on `New Folder`, enter the name `Repos` and click `Ok`.
2. In RStudio or RStudio Server select `File > New Project...`
3. Select `Version Control` and then `Git` (you might need to set up Git first if you are working on your own computer)
4. Copy `https://github.com/sdesabbata/granolarr.git` in the `Repository URL` field and select the Repos folder for the field `Create project as subdirectory of`, and click on `Create Project`.
5. Have a look around
6. Click on the project name `granolarr` on the top-left of the interface and select `Close Project` to close the project.

As granolarr is a public repository, you can clone it, edit it as you wish and push it to your own copy of the repository. However, contributing your edits to the original repository would require a few further steps. Check out the GitHub help pages if you are interested.

Chapter 6

Exploratory data analysis

Stefano De Sabbata

6.1 Introduction

This practical showcases an exploratory analysis of the distribution of people aged 20 to 24 in Leicester, using the `u011` variable from the 2011 Output Area Classification (2011OAC) dataset. Create a new R project for this practical session and create a new RMarkdown document to replicate the analysis in this document.

Once the document is set up, start by adding the first R code snipped including the code below, which is loads the 2011OAC dataset and the libraries used for the practical session.

```
library(tidyverse)
library(knitr)
leicester_2011OAC <- read_csv("2011_OAC_Raw_uVariables_Leicester.csv")
```

6.2 GGlot2 recap

As seen in the practical session 401, the `ggplot2` library is part of the Tidyverse, and it offers a series of functions for creating graphics **declaratively**, based on the concepts outlined in the Grammar of Graphics. While the `dplyr` library offers functionalities that cover *data manipulation* and *variable transformations*, the `ggplot2` library offers functionalities that allow to specify elements, define guides, and apply scale and coordinate system transformations.

- **Marks** can be specified in `ggplot2` using the `geom_` functions.
- The mapping of variables (table columns) to **visual variables** can be specified in `ggplot2` using the `aes` element.
- Furthermore, the `ggplot2` library:
  - automatically adds all necessary **guides** using default table column names, and additional functions can be used to overwrite the defaults;
  - provides a wide range of `scale_` functions that can be used to control the **scales** of all visual variables;
  - provides a series of `coord_` fucntions that allow transforming the **coordinate system**.

Check out the `ggplot2` reference for all the details about the functions and options discussed below.

6.3 Data visualisation

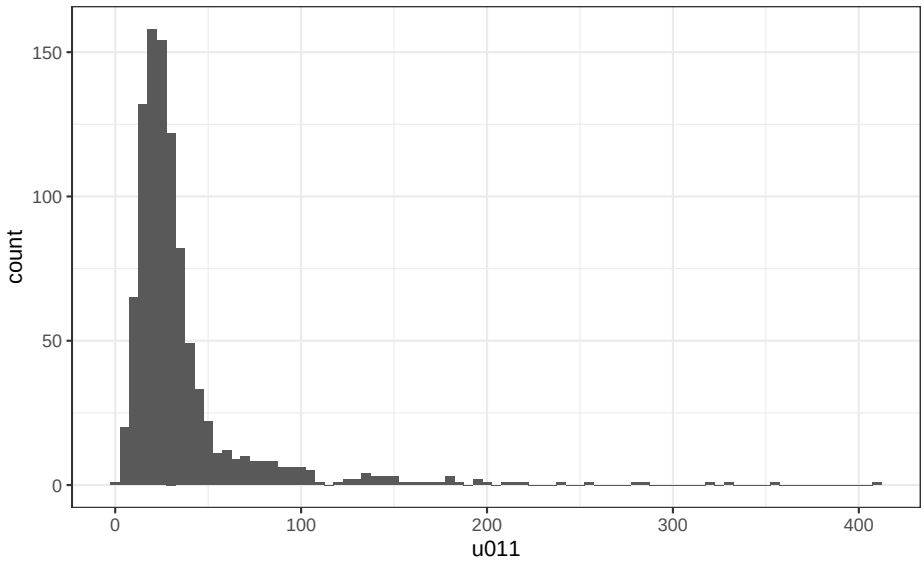
6.3.1 Distributions

We start the analysis with a simple histogram, to explore the distribution of the variable `u011`. RMarkdown allows specifying the height (as well as the width) of the figure as an option for the R snipped, as shown in the example typed out in plain text below.

````
```{r, echo=TRUE, message=FALSE, warning=FALSE, fig.height = 4}
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = u011
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```
````

The snipped and barchart is included in output documents, as shown below.

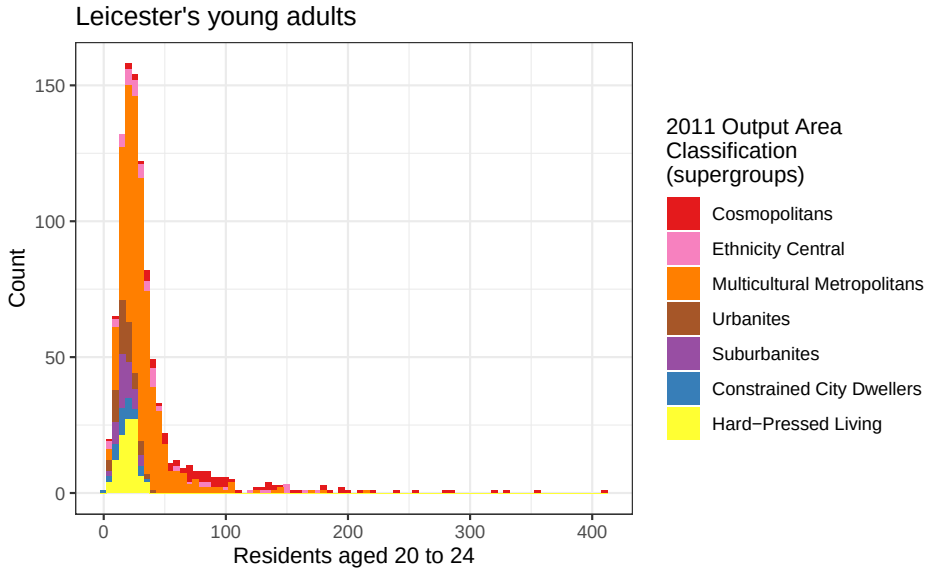
```
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = u011
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::theme_bw()
```



If we aim to explore how that portion of the population is distributed among the different supergroups of the 2011OAC, there are a number of charts that would allow us to visualise that relationship.

For instance, the barchart above can be enhanced through the use of the visual variable colour and the `fill` option. The graphic below uses a few options seen in the practical session 401 to create a stacked barchart, where sections of each bar are filled with the colour associated with a 2011OAC supergroup.

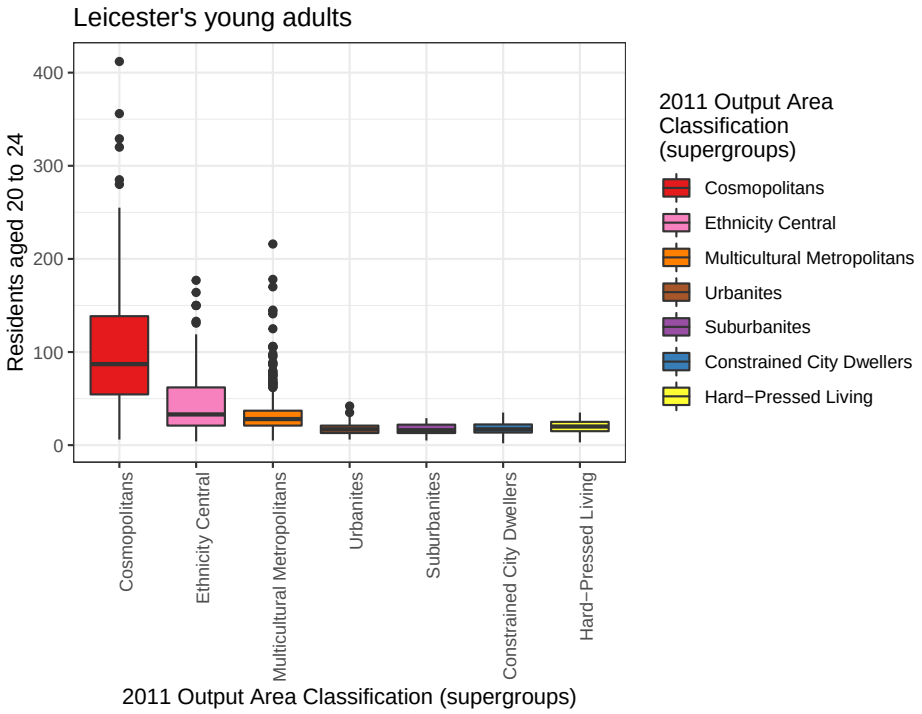
```
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = u011,
      fill = fct_reorder(supgrpname, supgrpcode)
    )
  ) +
  ggplot2::geom_histogram(binwidth = 5) +
  ggplot2::ggtitle("Leicester's young adults") +
  ggplot2::labs(
    fill = "2011 Output Area\nClassification\n(supergroups)"
  ) +
  ggplot2::xlab("Residents aged 20 to 24") +
  ggplot2::ylab("Count") +
  ggplot2::scale_fill_manual(
    values = c("#e41a1c", "#f781bf", "#ff7f00", "#a65628", "#984ea3", "#377eb8", "#ffff33")
  ) +
  ggplot2::theme_bw()
```



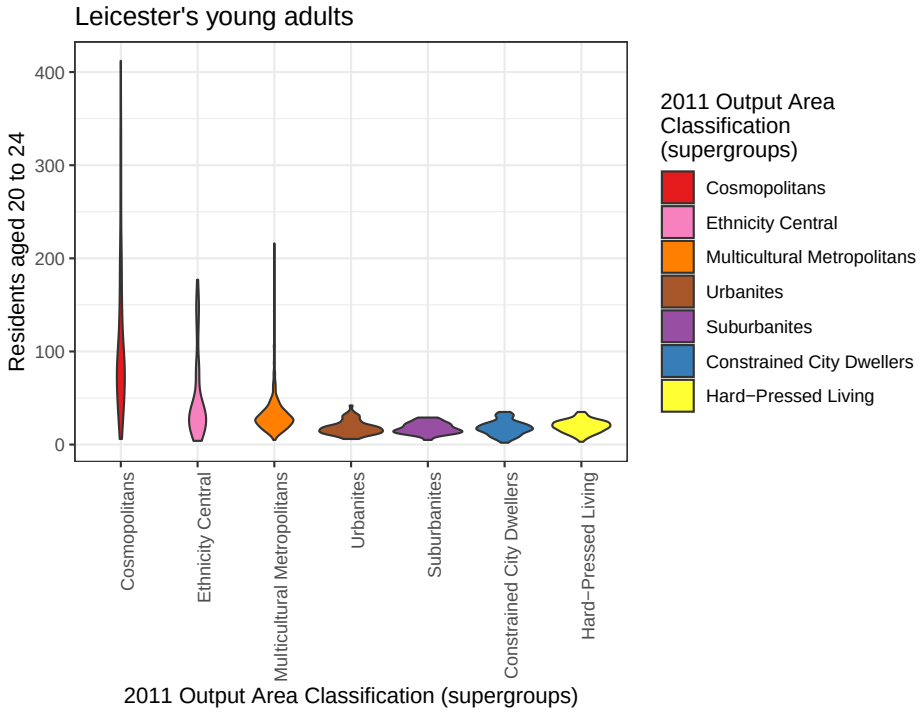
However, the graphic above is not extremely clear. A boxplot and a violin plot created from the same data are shown below. In both cases, the parameter `axis.text.x` of the function theme is set to `element_text(angle = 90, hjust = 1)` in order to orientate the labels on the x-axis vertically, as the supergroup names are rather long, and they would overlap one-another if set horizontally on the x-axis. In both cases, the option `fig.height` of the R snippet in RMarkdown should be set to a higher value (e.g., `5`) to allow for sufficient room for the supergroup names.

```
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = fct_reorder(supgrpname, supgrpcode),
      y = u011,
      fill = fct_reorder(supgrpname, supgrpcode)
    )
  ) +
  ggplot2::geom_boxplot() +
  ggtitle("Leicester's young adults") +
  ggplot2::labs(
    fill = "2011 Output Area\nClassification\n(supergroups)"
  ) +
  ggplot2::xlab("2011 Output Area Classification (supergroups)") +
  ggplot2::ylab("Residents aged 20 to 24") +
  ggplot2::scale_fill_manual(
    values = c("#e41a1c", "#f781bf", "#ff7f00", "#a65628", "#984ea3", "#377eb8", "#fff
  ) +
  ggplot2::theme_bw() +
```

```
  ggplot2::theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = fct_reorder(supgrpname, supgrpcode),
      y = u011,
      fill = fct_reorder(supgrpname, supgrpcode)
    )
  ) +
  ggplot2::geom_violin() +
  ggtitle("Leicester's young adults") +
  ggplot2::labs(
    fill = "2011 Output Area\nClassification\n(supergroups)"
  ) +
  ggplot2::xlab("2011 Output Area Classification (supergroups)") +
  ggplot2::ylab("Residents aged 20 to 24") +
  ggplot2::scale_fill_manual(
    values = c("#e41a1c", "#f781bf", "#ff7f00", "#a65628", "#984ea3", "#377eb8", "#ffff33")
  ) +
  ggplot2::theme_bw() +
  ggplot2::theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

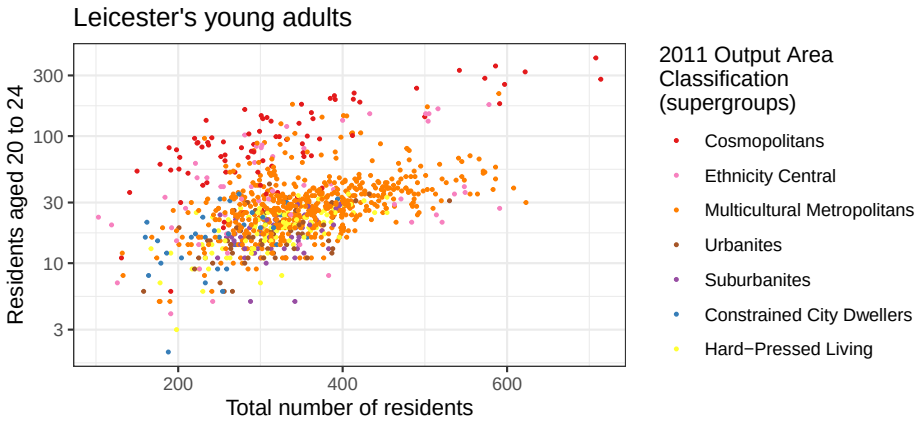


6.3.2 Relationships

The first barchart above seems to illustrate that the distribution might be skewed towards the left, with most values seemingly below 50. However, that tells only part of the story about how people aged 20 to 24 are distributed in Leicester. In fact, each Output Area (OA) has a different total population. So, a higher number of people aged 20 to 24 living in an OA might be simply due to the OA been more populous than others. Thus, the next step is to compare `u011` to `Total_Population`, for instance, through a scatterplot such as the one seen in the practical session 401, reported below.

```
leicester_2011OAC %>%
  ggplot2::ggplot(
    aes(
      x = Total_Population,
      y = u011,
      colour = fct_reorder(supgrpname, supgrpcode)
    )
  ) +
  ggplot2::geom_point(size = 0.5) +
  ggplot2::ggtitle("Leicester's young adults") +
  ggplot2::labs(
    colour = "2011 Output Area\nClassification\n(supergroups)"
```

```
  ) +
  ggplot2::xlab("Total number of residents") +
  ggplot2::ylab("Residents aged 20 to 24") +
  ggplot2::scale_y_log10() +
  ggplot2::scale_colour_brewer(palette = "Set1") +
  ggplot2::scale_colour_manual(
    values = c("#e41a1c", "#f781bf", "#ff7f00", "#a65628", "#984ea3", "#377eb8", "#ffff33")
  ) +
  ggplot2::theme_bw()
```



6.4 Exercise 304.1

Question 304.1.1: Which one of the boxplot or violin plot above do you think better illustrate the different distributions, and what do the two graphics say about the distribution of people aged 20 to 24 in Leicester? Write a short answer in your RMarkdown document (max 200 words).

Question 304.1.2: Create a jittered points plot (see `geom_jitter`) visualisation illustrating the same data shown in the boxplot and violin plot above.

Question 304.1.3: Create the code necessary to calculate a new column named `perc_age_20_to_24`, which is the percentage of people aged 20 to 24 (i.e., `u011`) over total population per OA `Total_Population`, and create a boxplot visualising the distribution of the variable per 2011OAC supergroup.

6.5 Exploratory statistics

The graphics above provide preliminary evidence that the distribution of people aged 20 to 24 might, in fact, be different in different 2011 supergroups. In the remainder of the practical session, we are going to explore that hypothesis further. First, load the necessary statistical libraries.

The code below calculates the percentage of people aged 20 to 24 (i.e., `u011`) over total population per OA, but it also recodes (see recode) the names of the 2011OAC supergroups to a shorter 2-letter version, which is useful for the tables presented further below.

Only the OA code, the recoded 2011OAC supergroup name, and the newly created `perc_age_20_to_24` are retained in the new table `leic_2011OAC_20to24`. Such a step is sometimes useful as stepping stone for further analysis and can make the code easier to read further down the line. Sometimes it is also a necessary step when interacting with certain libraries, which are not fully compatible with Tidyverse libraries, such as `leveneTest`.

```
leic_2011OAC_20to24 <- leicester_2011OAC %>%
  dplyr::mutate(
    perc_age_20_to_24 = (u011 / Total_Population) * 100,
    supgrpname = dplyr::recode(supgrpname,
      `Suburbanites` = "SU",
      `Cosmopolitans` = "CP",
      `Multicultural Metropolitans` = "MM",
      `Ethnicity Central` = "EC",
      `Constrained City Dwellers` = "CD",
      `Hard-Pressed Living` = "HP",
      `Urbanites` = "UR"
    )
  ) %>%
  dplyr::select(OA11CD, supgrpname, perc_age_20_to_24)

leic_2011OAC_20to24 %>%
  dplyr::slice_head(n = 5) %>%
  knitr::kable()
```

OA11CD	supgrpname	perc_age_20_to_24
E00069517	SU	4.153355
E00069514	CP	30.650155
E00169516	MM	12.316716
E00169048	MM	6.956522
E00169044	MM	6.211180

6.5.1 Descriptive statistics

The first step of any statistical analysis or modelling should be to explore the *"shape"* of the data involved, by looking at the descriptive statistics of all variables involved. The function `stat.desc` of the `pastecs` library provides three series of descriptive statistics.

- `base`:
  - `nbr.val`: overall number of values in the dataset;

  - `nbr.null`: number of `NULL` values – NULL is often returned by expressions and functions whose values are undefined;
  - `nbr.na`: number of `NA`s – missing value indicator;
- `desc`:
  - `min` (see also `min` function): **minimum** value in the dataset;
  - `max` (see also `max` function): **minimum** value in the dataset;
  - `range`: difference between `min` and `max` (different from `range()`);
  - `sum` (see also `sum` function): sum of the values in the dataset;
  - `median` (see also `median` function): **median**, that is the value separating the higher half from the lower half the values
  - `mean` (see also `mean` function): **arithmetic mean**, that is `sum` over the number of values not `NA`;
  - `SE.mean`: **standard error of the mean** – estimation of the variability of the mean calculated on different samples of the data (see also *central limit theorem*);
  - `CI.mean.0.95`: **95% confidence interval of the mean** – indicates that there is a 95% probability that the actual mean is within that distance from the sample mean;
  - `var`: **variance** (σ^2), it quantifies the amount of variation as the average of squared distances from the mean;
  - `std.dev`: **standard deviation** (σ), it quantifies the amount of variation as the square root of the variance;
  - `coef.var`: **variation coefficient** it quantifies the amount of variation as the standard deviation divided by the mean;
- `norm` (default is `FALSE`, use `norm = TRUE` to include it in the output):
  - `skewness`: **skewness** value indicates
    - positive: the distribution is skewed towards the left;
    - negative: the distribution is skewed towards the right;
  - `kurtosis`: **kurtosis** value indicates:
    - positive: heavy-tailed distribution;
    - negative: flat distribution;
  - `skew.2SE` and `kurt.2SE`: skewness and kurtosis divided by 2 standard errors. If greater than 1, the respective statistics is significant ($p < .05$);
  - `normtest.W`: test statistics for the **Shapiro–Wilk test** for normality;
  - `normtest.p`: significance for the **Shapiro–Wilk test** for normality.

The Shapiro–Wilk test compares the distribution of a variable with a normal distribution having the same mean and standard deviation. The null hypothesis of the Shapiro–Wilk test is that the sample is normally distributed, thus if `normtest.p` is lower than `0.01` (i.e., $p < .01$), the test indicates that the distribution is most probably not normal. The threshold to accept or reject a hypothesis is arbitrary and based on conventions, where $p < .01$ is the most commonly accepted threshold, or $p < .05$ for relatively small data sample (e.g., 30 cases).

The next step is thus to apply the `stat.desc` to the variable we are currently exploring (i.e., `perc_age_20_to_24`), including the `norm` section.

```
leic_2011OAC_20to24_stat_desc <- leic_2011OAC_20to24 %>%
  dplyr::select(perc_age_20_to_24) %>%
  pastecs::stat.desc(norm = TRUE)

leic_2011OAC_20to24_stat_desc %>%
  knitr::kable(digits = 3)
```

	perc_age_20_to_24
nbr.val	969.000
nbr.null	0.000
nbr.na	0.000
min	1.064
max	60.751
range	59.687
sum	10238.502
median	7.514
mean	10.566
SE.mean	0.304
CI.mean.0.95	0.596
var	89.386
std.dev	9.454
coef.var	0.895
skewness	2.710
skew.2SE	17.249
kurtosis	7.707
kurt.2SE	24.549
normtest.W	0.645
normtest.p	0.000

The table above tells us that all 969 OA in Leicester have a valid value for the variable `perc_age_20_to_24`, as no `NULL` nor `NA` value have been found. The values vary from about 1% to almost 61%, with an average value of 11% of the population in an OA aged between 20 and 24.

The short paragraph above is reporting on the values on the table, taking advantage of two features of RMarkdown. First, the output of the `stat.desc` function in the snippet further above is stored in the variable `leic_2011OAC_20to24_stat_desc`, which is then a valid variable for the rest of the document. Second, RMarkdown allows for in-line R snippets, that can also refer to variables defined in any snippet above the text. As such, the source of the paragraph above reads as below, with the in-line R snipped opened by a single grave accent (i.e., `` ` ``) followed by a lowercase `r` and closed by another single grave accent.

Having included all the code above into an RMarkdown document, copy the text below verbatim into the same RMarkdown document and make sure that you understand how the code in the in-line R snippets works.

```
The table above tells us that all `r "\u0060r leic_2011OAC_20to24_stat_desc[\"nbr.val\",
\"perc_age_20_to_24\"] %>% round(digits = 0)\u0060"` OA in Leicester have a valid
value for the variable `perc_age_20_to_24`, as no `r "\u0060NULL\u0060"` nor `r
"\u0060NA\u0060"` value have been found.The values vary from about `r "\u0060r
leic_2011OAC_20to24_stat_desc[\"min\", \"perc_age_20_to_24\"] %>% round(digits =
0)\u0060"`% to almost `r "\u0060r leic_2011OAC_20to24_stat_desc[\"max\",
\"perc_age_20_to_24\"] %>% round(digits = 0)\u0060"`%, with an average value of
`r "\u0060r leic_2011OAC_20to24_stat_desc[\"mean\", \"perc_age_20_to_24\"] %>%
round(digits = 0)\u0060"`% of the population in an OA aged between 20 and 24.
```

If the data described by statistics presented in the table above was a random sample of a population, the 95% confidence interval `CI.mean.0.95` would indicate that we can be 95% confident that the actual mean of the distribution is somewhere between 10.566 - 0.596 = 9.97% and 10.566 + 0.596 = 11.162%.

However, this is not a sample. Thus the statistical interpretation is not valid, in the same way that the `sum` values doesn't make sense, as it is the sum of a series of percentages.

Both `skew.2SE` and `kurt.2SE` are greater than `1`, which indicate that the `skewness` and `kurtosis` values are significant ($p < .05$). The `skewness` is positive, which indicates that the distribution is skewed towards the left (low values). The `kurtosis` is positive, which indicates that the distribution is heavy-tailed.

As such, `perc_age_20_to_24` having a heavy-tailed distribution skewed towards low values, it is not surprising that the `normtest.p` value indicates that the Shapiro–Wilk test is significant, which indicates that the distribution is not normal.

The code below present the output of the `shapiro.test` function, which only present the outcome of a Shapiro–Wilk test on the values provided as input. The output values are the same as the values reported by the `norm` section of `stat.desc`. Note that the `shapiro.test` function require the argument to be a numeric vector. Thus the `pull` function must be used to extract the `perc_age_20_to_24` column from `leic_2011OAC_20to24` as a vector, whereas using `select` with a single column name as the argument would produce as output a table with a single column.

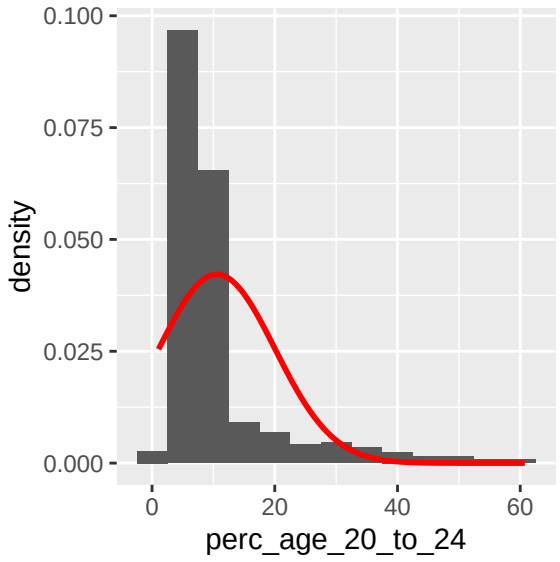
```
leic_2011OAC_20to24 %>%
  dplyr::pull(perc_age_20_to_24) %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  .
## W = 0.64491, p-value < 2.2e-16
```

The two code snippets below can be used to visualise a density-based histogram including the shape of a normal distribution having the same mean and standard deviation, and a Q-Q plot, to visually confirm the fact that `perc_age_20_to_24` is not normally distributed.

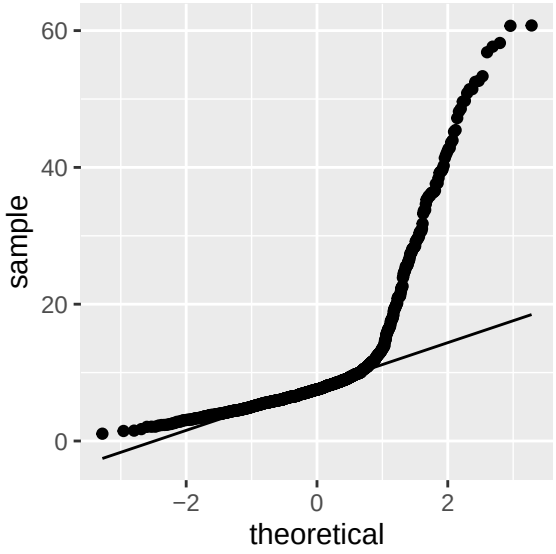
```
leic_2011OAC_20to24 %>%
  ggplot2::ggplot(
    aes(
      x = perc_age_20_to_24
    )
  ) +
  ggplot2::geom_histogram(
    aes(
      y =..density..
    ),
    binwidth = 5
  ) +
  ggplot2::stat_function(
    fun = dnorm,
    args = list(
      mean = leic_2011OAC_20to24 %>% pull(perc_age_20_to_24) %>% mean(),
      sd = leic_2011OAC_20to24 %>% pull(perc_age_20_to_24) %>% sd()
    ),
    colour = "red", size = 1
  )
```



A Q-Q plot in R can be created using a variety of functions. In the example below, the plot is created using the `stat_qq` and `stat_qq_line` functions of the `ggplot2` library. Note that the `perc_age_20_to_24` variable is mapped to a particular option of `aes` that is `sample`.

If `perc_age_20_to_24` had been normally distributed, the dots in the Q-Q plot would be distributed straight on the line included in the plot.

```
leic_2011OAC_20to24 %>%
  ggplot2::ggplot(
    aes(
      sample = perc_age_20_to_24
    )
  ) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line()
```



6.6 Exercise 304.2

Create a new RMarkdown document, and add the code necessary to recreate the table `leic_2011OAC_20to24` used in the example above. Use the code below to re-shape the table `leic_2011OAC_20to24` by pivoting the `perc_age_20_to_24` column wider into multiple columns using `supgrpname` as new column names.

```
leic_2011OAC_20to24_supgrp <- leic_2011OAC_20to24 %>%
  tidyr::pivot_wider(
    names_from = supgrpname,
    values_from = perc_age_20_to_24
```

```
  )
```

That manipulation creates one column per supergroup, containing the `perc_age_20_to_24` if the OA is part of that supergroup, or an `NA` value if the OA is not part of the supergroup. The transformation is illustrated in the two tables below. The first shows an extract from the original `leic_2011OAC_20to24` dataset, followed by the wide version `leic_2011OAC_20to24_supgrp`.

```
leic_2011OAC_20to24 %>%
  dplyr::slice_min(OA11CD, n = 10) %>%
  knitr::kable(digits = 3)
```

OA11CD	supgrpname	perc_age_20_to_24
E00068657	HP	6.053
E00068658	MM	6.964
E00068659	MM	8.383
E00068660	MM	4.643
E00068661	MM	10.625
E00068662	MM	8.284
E00068663	MM	8.357
E00068664	MM	3.597
E00068665	MM	7.068
E00068666	MM	5.864

```
leic_2011OAC_20to24_supgrp %>%
  dplyr::slice_min(OA11CD, n = 10) %>%
  knitr::kable(digits = 3)
```

OA11CD	SU	CP	MM	EC	CD	HP	UR
E00068657	NA	NA	NA	NA	NA	6.053	NA
E00068658	NA	NA	6.964	NA	NA	NA	NA
E00068659	NA	NA	8.383	NA	NA	NA	NA
E00068660	NA	NA	4.643	NA	NA	NA	NA
E00068661	NA	NA	10.625	NA	NA	NA	NA
E00068662	NA	NA	8.284	NA	NA	NA	NA
E00068663	NA	NA	8.357	NA	NA	NA	NA
E00068664	NA	NA	3.597	NA	NA	NA	NA
E00068665	NA	NA	7.068	NA	NA	NA	NA
E00068666	NA	NA	5.864	NA	NA	NA	NA

Question 304.2.1: The code below uses the newly created `leic_2011OAC_20to24_supgrp` table to calculate the descriptive statistics calculated for the variable `leic_2011OAC_20to24` for each supergroup. Is `leic_2011OAC_20to24` normally distributed in any of the subgroups? If yes, which supergroups and based on which values do you justify that claim? (Write up to 200 words)

```
leic_2011OAC_20to24_supgrp %>%
  dplyr::select(-OA11CD) %>%
  pastecs::stat.desc(norm = TRUE) %>%
  knitr::kable(digits = 3)
```

	SU	CP	MM	EC	CD	HP	UR
nbr.val	54.000	83.000	573.000	57.000	36.000	101.000	65.000
nbr.null	0.000	0.000	0.000	0.000	0.000	0.000	0.000
nbr.na	915.000	886.000	396.000	912.000	933.000	868.000	904.000
min	1.462	3.141	2.490	2.066	1.064	1.515	2.256
max	9.562	60.751	52.507	36.299	12.963	11.261	13.505
range	8.100	57.609	50.018	34.233	11.899	9.746	11.249
sum	295.867	2646.551	5214.286	838.415	252.108	619.266	372.010
median	5.476	30.457	7.880	10.881	6.854	6.053	5.380
mean	5.479	31.886	9.100	14.709	7.003	6.131	5.723
SE.mean	0.233	1.574	0.230	1.373	0.471	0.172	0.264
CI.mean.0.95	0.467	3.131	0.452	2.751	0.956	0.341	0.528
var	2.929	205.556	30.285	107.523	7.983	2.980	4.545
std.dev	1.712	14.337	5.503	10.369	2.825	1.726	2.132
coef.var	0.312	0.450	0.605	0.705	0.403	0.282	0.372
skewness	0.005	0.067	3.320	0.633	0.322	0.124	1.042
skew.2SE	0.008	0.127	16.266	1.001	0.410	0.258	1.753
kurtosis	-0.391	-0.825	15.143	-1.009	-0.142	0.220	1.441
kurt.2SE	-0.306	-0.789	37.156	-0.810	-0.093	0.231	1.229
normtest.W	0.991	0.980	0.684	0.889	0.965	0.993	0.937
normtest.p	0.954	0.239	0.000	0.000	0.310	0.886	0.002

Question 304.2.2: Write the code necessary to test again the normality of `leic_2011OAC_20to24` for the supergroups where the analysis conducted for Question 304.2.1 indicated they are normal, using the function `shapiro.test`, and draw the respective Q-Q plot.

Question 304.2.3: Observe the output of the Levene's test executed below. What does the result tell you about the variance of `perc_age_20_to_24` in supergroups?

```
car::leveneTest(leic_2011OAC_20to24$perc_age_20_to_24, leic_2011OAC_20to24$supgrpname)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   6  62.011 < 2.2e-16 ***
##       962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 7

Comparing data

Stefano De Sabbata

7.1 Introduction

The first part of this practical guides you through the ANOVA (analysis of variance) and regression analysis seen in the lecture, the last part showcases a multiple regression analysis. Create a new R project for this practical session and create a new RMarkdown document to replicate the analysis in this document and a separate RMarkdown document to work on the exercises.

```
library(tidyverse)
library(magrittr)
library(knitr)
```

As many of the functions used in the analyses below are part of the oldest libraries developed for R, they have not been developed to be easily compatible with the Tidyverse and the `%>%` operator. Fortunately, the `magrittr` library (loaded above) does not only define the `%>%` operator seen so far, but also the exposition pipe operator `%$%`, which exposes the columns of the data.frame on the left of the operator to the expression on the right of the operator. That is, `%$%` allows to refer to the column of the data.frame directly in the subsequent expression. As such, the lines below expose the column `Petal.Length` of the data.frame `iris` and to pass it on to the `mean` function using different approaches, but they are all equivalent in their outcome.

```
# Classic R approach
mean(iris$Petal.Length)
```

```
## [1] 3.758
```

```
# Using %>% pipe
iris$Petal.Length %>%
  mean()
```

```
## [1] 3.758
```

```
# Using %>% pipe and %$% exposition pipe
iris %$% Petal.Length %>%
  mean()
```

```
## [1] 3.758
```

7.2 ANOVA

The ANOVA (analysis of variance) tests whether the values of a variable (e.g., length of the petal) are on average different for different groups (e.g., different species of iris). ANOVA has been developed as a generalised version of the t-test, which has the same objective but allows to test only two groups.

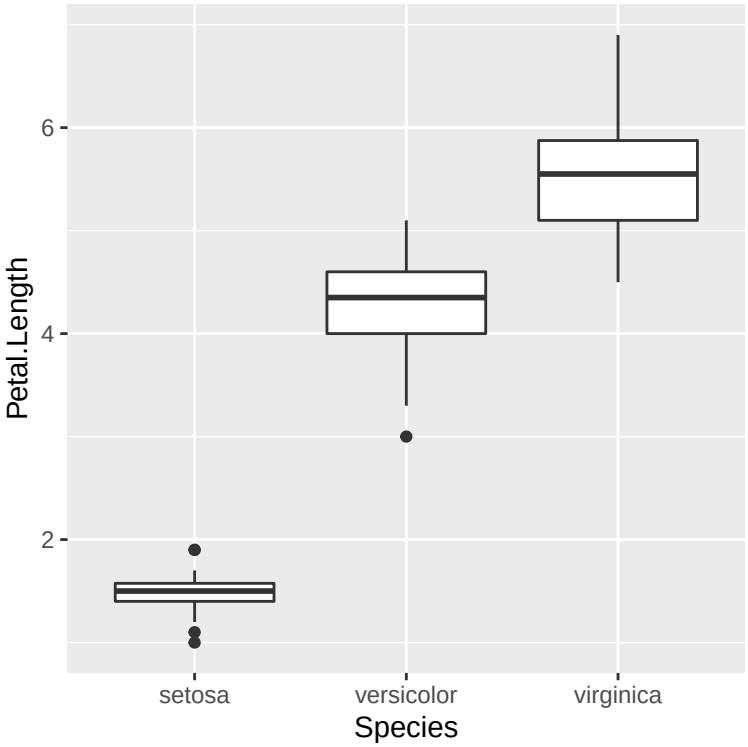
The ANOVA test has the following assumptions:

- normally distributed values in groups
  - especially if groups have different sizes
- homogeneity of variance of values in groups
  - if groups have different sizes
- independence of groups

7.2.1 Example

The example seen in the lecture illustrates how ANOVA can be used to verify that the three different species of iris in the `iris` dataset have different petal length.

```
iris %>%
  ggplot2::ggplot(
    aes(
      x = Species,
      y = Petal.Length
    )
  ) +
  ggplot2::geom_boxplot()
```



ANOVA is considered a robust test, thus, as the groups are of the same size, there is no need to test for the homogeneity of variance. Furthermore, the groups come from different species of flowers, so there is no need to test the independence of the values. The only assumption that needs testing is whether the values in the three groups are normally distributed. As there are 50 flowers per species, we can set the significance threshold to `0.05`.

The three Shapiro–Wilk tests below are all not significant, which indicates that all three groups have normally distributed values.

```
iris %>% dplyr::filter(Species == "setosa") %>% dplyr::pull(Petal.Length) %>% stats::shapiro.test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.95498, p-value = 0.05481
```

```
iris %>% dplyr::filter(Species == "versicolor") %>% dplyr::pull(Petal.Length) %>% stats::shapiro.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
```

```
## W = 0.966, p-value = 0.1585
```

```
iris %>% dplyr::filter(Species == "virginica") %>% dplyr::pull(Petal.Length) %>% stats
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.96219, p-value = 0.1098
```

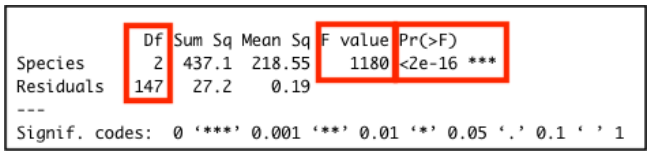
We can thus conduct the ANOVA test using the function `aov`, and the function `summary` to obtain the summary of the results of the test.

```
# Classic R coding approach (not using %$%)
# iris_anova <- aov(Petal.Length ~ Species, data = iris)
# summary(iris_anova)

iris %$%
  stats::aov(Petal.Length ~ Species) %>%
  summary()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference is significant $\mathrm{F}(2, 147) = 1180.16$, $p < .01$.

The image below highlights the important values in the output: the significance value `Pr(>F)`; the F-statistic value `F value`; and the two degrees of freedom values for the F-statistic in the `Df` column.



7.3 Exercise 314.1

Question 314.1.1: Load the `2011_OAC_Raw_uVariables_Leicester.csv` dataset. Check whether the values of mean age (`u020`) are normally distributed, and whether they can be transformed to a normally distributed set using logarithmic or inverse hyperbolic sine functions.

Question 314.1.2: Check whether the values of mean age (`u020`) are normally distributed when looking at the different 2011OAC supergroups separately. Check whether they can be transformed to a normally distributed set

using logarithmic or inverse hyperbolic sine functions.

Question 314.1.3: Is the distribution of mean age (`u020`) different in different 2011OAC supergroups in Leicester?

7.4 Correlation

The term **correlation** is used to refer to a series of a standardised measures of covariance, which can be used to statistically assess whether two variables are related or not.

Furthermore, if two variables are related, such measures can identify whether they are:

- positively related:
  - entities with *high values* in one tend to have *high values* in the other;
  - entities with *low values* in one tend to have *low values* in the other;
- negatively:
  - entities with *high values* in one tend to have *low values* in the other;
  - entities with *low values* in one tend to have *high values* in the other.

Correlation can be calculated in many ways, but there are three approaches which are by far the most common. They all start from the null hypothesis that there is no relationship between the variables. Thus, if the p-value is above a pre-defined significance threshold, the null hypothesis is rejected, and the conclusion is that there is a relationship between the two variables.

If the test is significant is the case:

- a **positive** correlation value indicates a positive relationship;
- a **negative** correlation value indicates a negative relationship;
- the **square** of the correlation value can be taken as an indication of the percentage of shared variance between the two variables.

However, each one has different assumptions about the variables' distribution and thus implements the same general ideal measure in a different way:

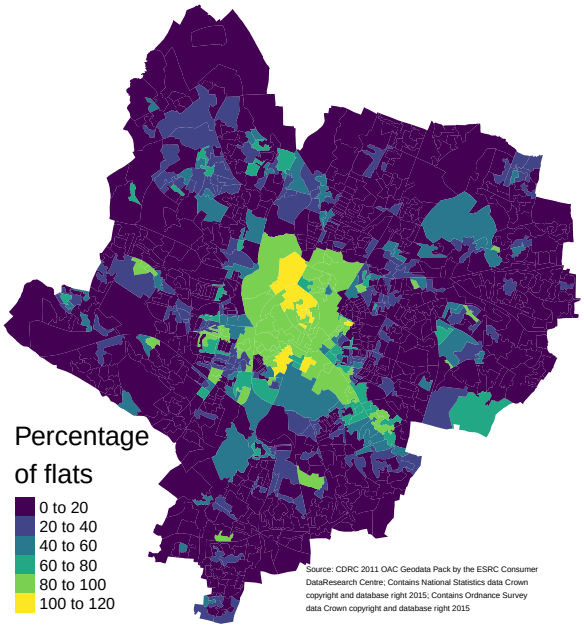
- if two variables are **normally distributed**:
  - *Pearson's r*;
- if two variables are **not normally distributed**:
  - if there are **no ties among values**:
    - *Spearman's rho*;
  - if there are **ties among values**:
    - *Kendall's tau*.

7.4.1 Example

When studying how people live in cities, a number of questions might arise about where the live and how the move around the city. For instance, looking

at a map of Leicester, it is clear that (has in many English cities) there seems to be a very high concentration of flats in the city centre. At the same time, there seems to be almost no flats at all in the suburbs. This might led us to ask: *"do households living in flats (and thus mostly in the city centre) own the same amount of cars as households living in the city center?"*

That could be due to many reasons. As the suburbs in England are largely residential, whereas most working places are located in the city centre. As such people living in flats might be more likely to walk or cycle to work, or commute using public transportation within the city or to other cities. City centres usually afford less spaces for parking. Many flats are rented to students, who might be less likely to own a car. The list could continue, but these are still hypothesis based on a certain (probably biased) view of the city. Can we use data analysis to explore whether there is any ground to such an hypothesis?



The dataset used to create the 2011 Output Area Classification (2011OAC) contains two variables that might help explore this issue. These data are not very current anymore, and they are not they values we might collect if we were to conduct a fresh survey for this specific study. However, they can still provide some insight.

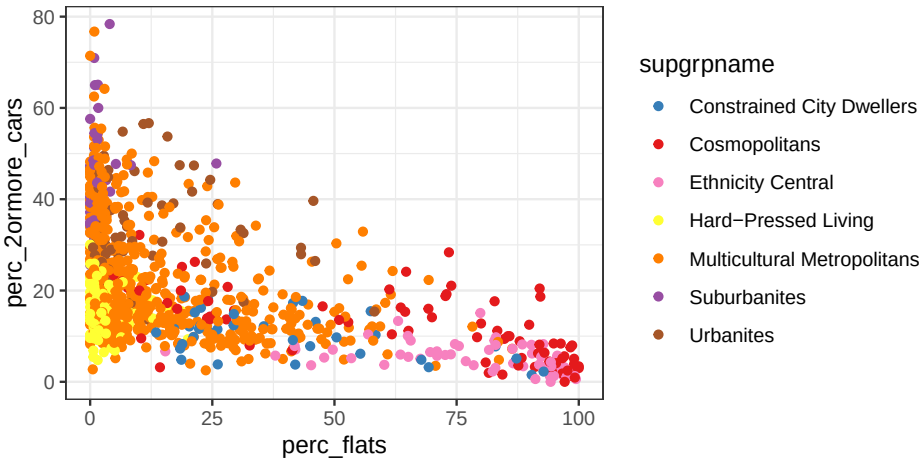
- `u089`: count of flats per Output Area (OA). The statistical unit for this variable is `Household_Spaces`. As OAs vary in size and composition, we can use `Total_Household_Spaces` to calculate the percentage of flats per OA, which is a more stable measure.
  - `perc_flats = (u089 / Total_Household_Spaces) * 100`
- `u118`: 2 or more cars or vans in household. The statistical unit for this

variable is `Household`. As OAs vary in size and composition, we can use `Total_Household_Spaces` to calculate the percentage of households per OA with 2 or more cars or vans, which is a more stable measure.
  - `perc_2ormore_cars = (u118 / Total_Households) * 100`

The process of transforming variables to be within a certain range (such as a percentage, thus using a `[0..100]` range, or a `[0..1]` range) is commonly referred to as **normalisation**. The process of transforming a variable to have mean zero and standard deviation one (z-scores) is commonly referred to as **standardisation**. However, note that these terms are sometime used interchangably.

```
flats_and_cars <-
  leicester_2011OAC %>%
  dplyr::mutate(
    perc_flats = (u089 / Total_Household_Spaces) * 100,
    perc_2ormore_cars = (u118 / Total_Households) * 100
  ) %>%
  dplyr::select(
    OA11CD, supgrpname, supgrpcode,
    perc_flats, perc_2ormore_cars
  )
```

Plotting the two variables together in a scatterplot reveals a pattern. Indeed, a very low percentage of households living in flats own two or more cars. However, the proportion of households owning two or more cars who live in the suburbs seem to span almost throughout the whole range, from zero to 80%. That seems to indicate some level of negative relationship, but the picture is clearly far less clear-cut as we might have initially assumed. The initial assumption about car ownership for households living in flats seems to hold, but we probably didn't consider the situation in the suburbs with sufficient care.



The first step in establishing whether there is a relationship between the two

variables is to assess whether they are normally distributed, and thus which correlation test we should use for the analysis. The scatterplot already seem to suggest that the variables are rather skewed.

As there are 969 OAs in Leicester, we can set the significance threshold to `0.01`. The results of the `stats::shapiro.test` functions below show that neither of the two variables are normally distributed. Transforming the variables using the *inverse hyperbolic sine* still does not result in normally distributed variables. Thus, we should discard *Pearson's r* as an option to explore the correlation between the two variables.

```
flats_and_cars %>%
  dplyr::select(perc_flats, perc_2ormore_cars) %>%
  dplyr::mutate(
    ihs_perc_flats = asinh(perc_flats),
    ihs_perc_2omcars = asinh(perc_2ormore_cars)
  ) %>%
  pastecs::stat.desc(basic = FALSE, desc = FALSE, norm = TRUE) %>%
  knitr::kable()
```

	perc_flats	perc_2ormore_cars	ihs_perc_flats	ihs_perc_2omcars
skewness	1.5621906	0.9075026	-0.0927406	-0.9460022
skew.2SE	9.9417094	5.7753049	-0.5901967	-6.0203149
kurtosis	1.3282688	0.4588571	-1.1009004	1.6988166
kurt.2SE	4.2308489	1.4615680	-3.5066270	5.4111309
normtest.W	0.7443821	0.9328442	0.9572430	0.9514757
normtest.p	0.0000000	0.0000000	0.0000000	0.0000000

The next step is to assess whether there are ties among the values in the two variables. The code below fist counts the number of cases per value. Then it counts the number of values for which the number of cases is greater than one.

```
ties_perc_flats <-
  flats_and_cars %>%
  dplyr::count(perc_flats) %>%
  dplyr::filter(n > 1) %>%
  # Specify wt = n() to count rows
  # otherwise n is taken as weight
  dplyr::count(wt = n()) %>%
  dplyr::pull(n)

ties_perc_2ormore_cars <-
  flats_and_cars %>%
  dplyr::count(perc_2ormore_cars) %>%
  dplyr::filter(n > 1) %>%
  # Specify wt = n() to count rows
  # otherwise n is taken as weight
  dplyr::count(wt = n()) %>%
```

```
  dplyr::pull(n)
```

The variable `perc_flats` has 127 values with ties and `perc_2ormore_cars` has 115 values with ties. As such, using *Spearman's rho* is not advisable and *Kendall's tau* should be used. As above, we can set the significance threshold to `0.01`.

Finally, we can run the `stats::cor.test` function to assess the relationship between the two variables. The code below saves the results of the test to a variable. This afford to subsequent actions. First, we can show the full results by simply invoking the name of the variable (term used in the programming-related meaning here) in the final line of the code. Second, we can extract and square the estimate value in RMarkdwon in the following paragraph, to show the percentage of shared variace.

```
flats_and_cars_corKendall <-
  flats_and_cars %$%
  stats::cor.test(
    perc_flats, perc_2ormore_cars,
    method = "kendall"
  )

flats_and_cars_corKendall
```

```
##
##  Kendall's rank correlation tau
##
## data:  perc_flats and perc_2ormore_cars
## z = -19.026, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.4094335
```

```
The percentage of flats and the percentage of households
owning 2 or more cars or vans per OA in the city of Leicester
are negative related, as the relationship is significant
(`p-value < 0.01`) and the correlation value is negative
(`tau =` -0.41). The two variables share 16.8% of variance. We can thus conclude
that there is significant but very weak relationship
between the two variables.
```

The percentage of flats and the percentage of households owning 2 or more cars or vans per OA in the city of Leicester are negative related, as the relationship is significant (`p-value < 0.01`) and the correlation value is negative (`tau =` -0.41). The two variables share 16.8% of variance. We can thus conclude that there is significant but very weak relationship between the two variables.

7.5 Exercise 314.2

Question 314.2.1: As mentioned above, when discussing movement in cities, there is an assumption that people living in the city centre live in flats and work or cycle to work, whereas people living in the suburbs live in whole houses and commute via car. Study the correlation between the presence of flats (`u089`) and people commuting to work on foot, bicycle or other similar means (`u122`) in the same OAs. Consider whether the values might need to be normalised or otherwised transformed before starting the testing procedure.

Question 314.2.2: Another interesting issue to explore is the relationship between car ownership and the use of public transport. Study the correlation between the presence of households owning 2 or more cars or vans (`u118`) and people commuting to work via public transport (`u120`) or on foot, bicycle or other similar means (`u122`) in the same OAs. Consider whether the values might need to be normalised or otherwised transformed before starting the testing procedure.

Chapter 8

Regression analysis

Stefano De Sabbata

8.1 Simple regression

The simple regression analysis is a supervised machine learning approach to creating a model able to predict the value of one outcome variable Y based on one predictor variable X_1, by estimating the intercept b_0 and coefficient (slope) b_1, and accounting for a reasonable amount of error ϵ.

$$Y_i = (b_0 + b_1 * X_{i1}) + \epsilon_i$$

Least squares is the most commonly used approach to generate a regression model. This model fits a line to minimise the squared values of the **residuals** (errors), which are calculated as the squared difference between observed values the values predicted by the model.

$$redidual = \sum (observed - model)^2$$

A model is considered **robust** if the residuals do not show particular trends, which would indicate that *"something"* is interfering with the model. In particular, the assumption of the regression model are:

- **linearity:** the relationship is actually linear;
- **normality** of residuals: standard residuals are normally distributed with mean `0`;

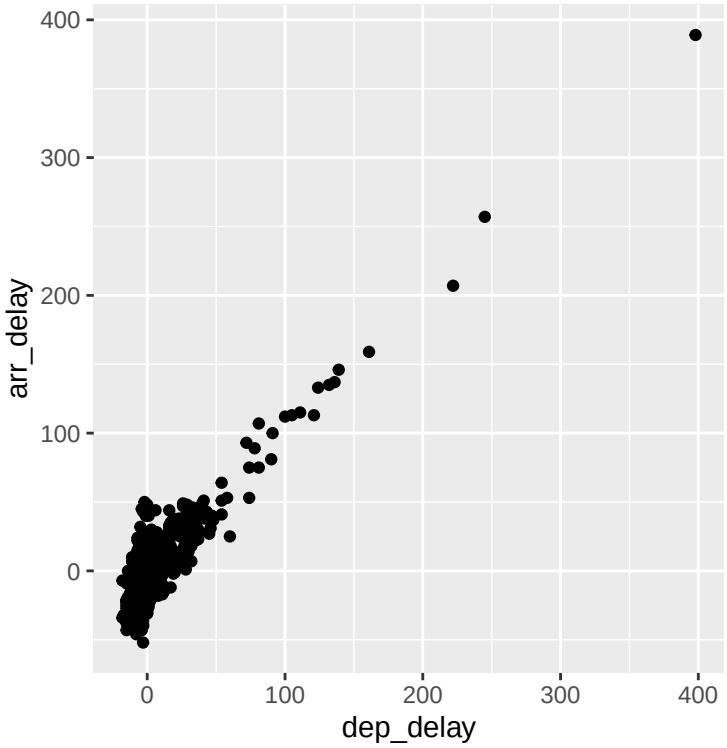
- **homoscedasticity** of residuals: at each level of the predictor variable(s) the variance of the standard residuals should be the same (*homo-scedasticity*) rather than different (*hetero-scedasticity*);
- **independence** of residuals: adjacent standard residuals are not correlated.

8.1.1 Example

The example that we have seen in the lecture illustrated how simple regression can be used to create a model to predict the arrival delay based on the departure delay of a flight, based on the data available in the `nycflights13` dataset for the flight on November 20th, 2013. The scatterplot below seems to indicate that the relationship is indeed linear.

$$arr\_delay_i = (Intercept + Coefficient_{dep\_delay} * dep\_delay_{i1}) + \epsilon_i$$

```
# Load the library
library(nycflights13)

# November 20th, 2013
flights_nov_20 <- nycflights13::flights %>%
  dplyr::filter(!is.na(dep_delay), !is.na(arr_delay), month == 11, day ==20)
```



The code below generates the model using the function `lm`, and the function `summary` to obtain the summary of the results of the test. The model and summary are saved in the variables `delay_model` and `delay_model_summary`, respectively, for further use below. The variable `delay_model_summary` can then be called directly to visualise the result of the test.

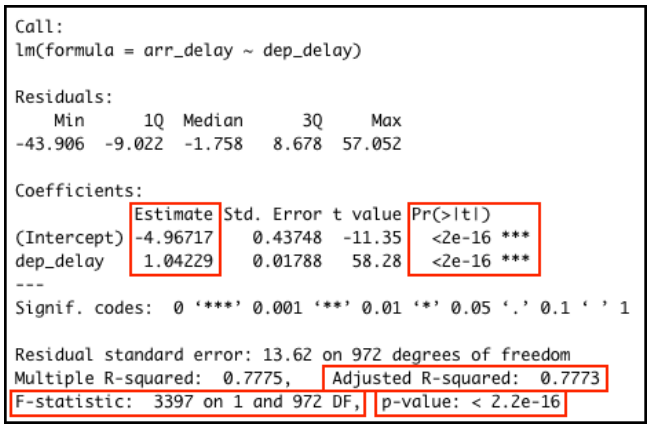
```
# Classic R coding version
# delay_model <- lm(arr_delay ~ dep_delay, data = flights_nov_20)
# delay_model_summary <- summary(delay_model)

# Load magrittr library to use %$%
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
delay_model <- flights_nov_20 %$%
  lm(arr_delay ~ dep_delay)

delay_model_summary <- delay_model %>%
  summary()

delay_model_summary
```

```
##
## Call:
## lm(formula = arr_delay ~ dep_delay)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.906  -9.022  -1.758   8.678  57.052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.96717    0.43748  -11.35   <2e-16 ***
## dep_delay    1.04229    0.01788   58.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.62 on 972 degrees of freedom
## Multiple R-squared:  0.7775,Adjusted R-squared:  0.7773
## F-statistic:  3397 on 1 and 972 DF,  p-value: < 2.2e-16
```

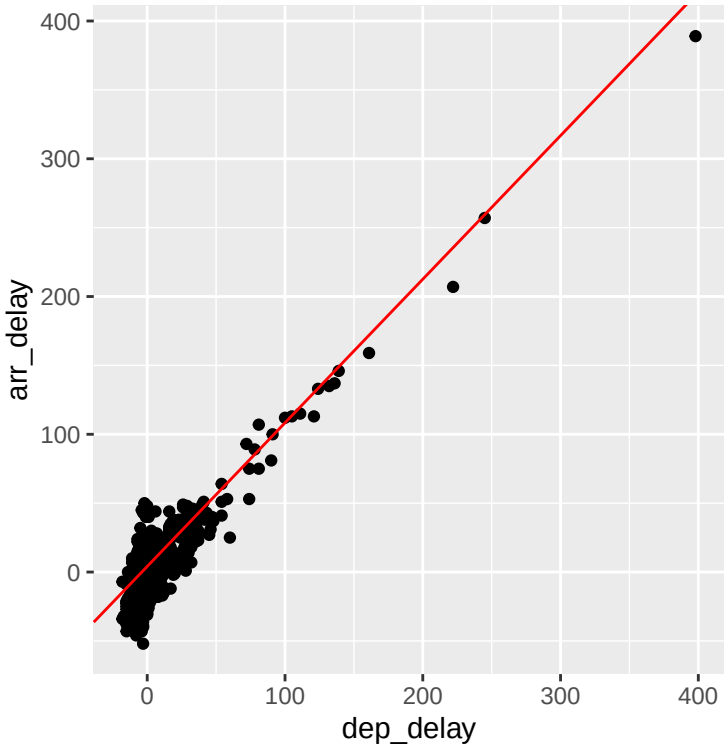
The image below highlights the important values in the output: the adjusted R^2 value; the model significance value `p-value` and the related F-statistic information `F-statistic`; the intercept and `dep_delay` coefficient estimates in the `Estimate` column and the related significance values of in the column `Pr(>|t|)`.



The output indicates:

- **p-value: < 2.2e-16**: $p < .001$ the model is significant;
  - derived by comparing the calulated **F-statistic** value to F distribution 3396.74 having specified degrees of freedom (1, 972);
  - Report as: $F(1, 972) = 3396.74$
- **Adjusted R-squared: 0.7773**: the departure delay can account for 77.73% of the arrival delay;
- **Coefficients**:
  - Intercept estimate -4.9672 is significant;
  - `dep_delay` coefficient (slope) estimate 1.0423 is significant.

```
flights_nov_20 %>%
  ggplot2::ggplot(aes(x = dep_delay, y = arr_delay)) +
  ggplot2::geom_point() + ggplot2::coord_fixed(ratio = 1) +
  ggplot2::geom_abline(intercept = 4.0943, slope = 1.04229, color="red")
```



8.1.2 Checking assumptions

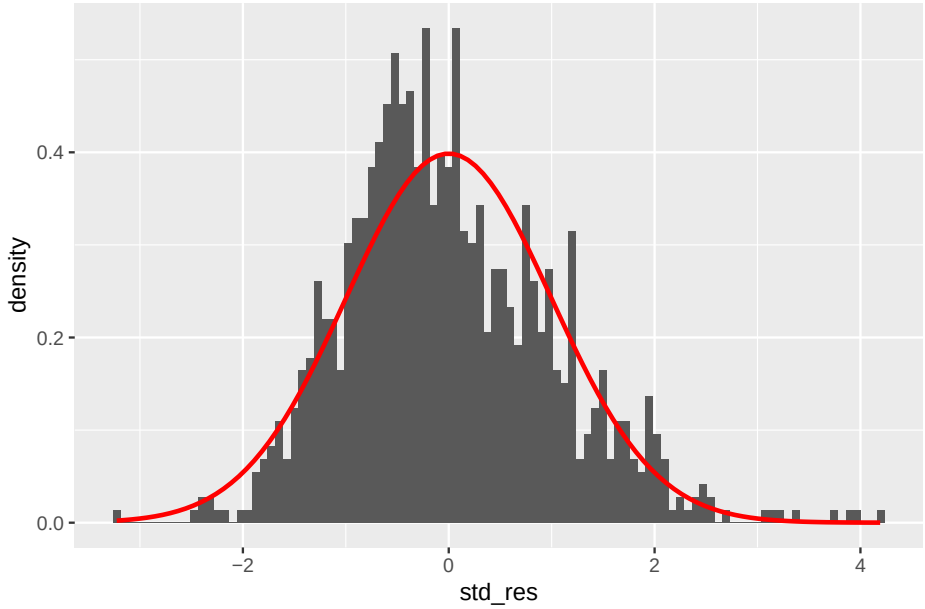
8.1.2.1 Normality

The Shapiro-Wilk test can be used to check for the normality of standard residuals. The test should be not significant for robust models. In the example

below, the standard residuals are *not* normally distributed. However, the plot further below does show that the distribution of the residuals is not far away from a normal distribution.

```
delay_model %>%
  stats::rstandard() %>%
  stats::shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.98231, p-value = 1.73e-09
```



8.1.2.2 Homoscedasticity

The Breusch-Pagan test can be used to check for the homoscedasticity of standard residuals. The test should be not significant for robust models. In the example below, the standard residuals are homoscedastic.

```
library(lmtest)

delay_model %>%
  lmtest::bptest()
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  .
## BP = 0.017316, df = 1, p-value = 0.8953
```

8.1.2.3 Independence

The Durbin-Watson test can be used to check for the independence of residuals. The test should be statistic should be close to 2 (between 1 and 3) and not significant for robust models. In the example below, the standard residuals might not be completely independent. Note, however, that the result depends on the order of the data.

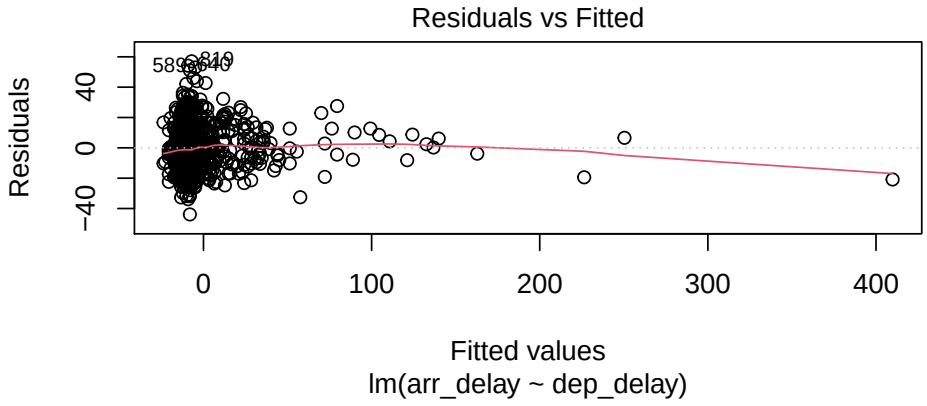
```
# Also part of the library lmtest
delay_model %>%
  lmtest::dwtest()
```

```
##
##  Durbin-Watson test
##
## data:  .
## DW = 1.8731, p-value = 0.02358
## alternative hypothesis: true autocorrelation is greater than 0
```

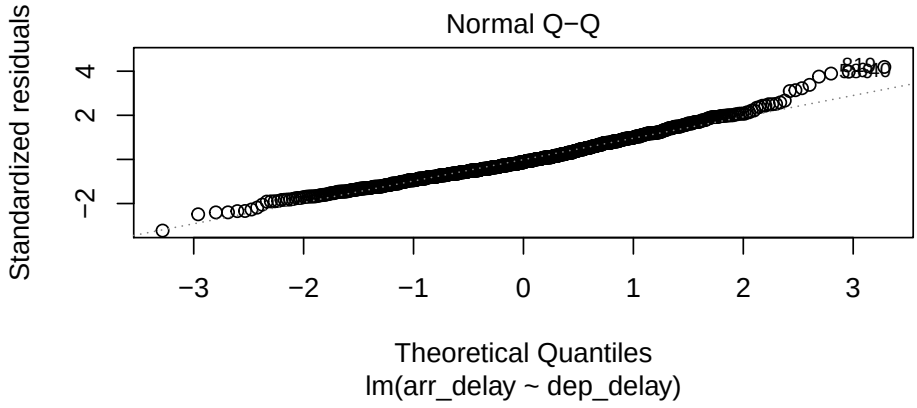
8.1.2.4 Plots

The `plot.lm` function can be used to further explore the residuals visuallly. Usage is illustrated below. The *Residuals vs Fitted* and *Scale-Location* plot provide an insight into the homoscedasticity of the residuals, the *Normal Q-Q* plot provides an illustration of the normality of the residuals, and the *Residuals vs Leverage* can be useful to identify exceptional cases (e.g., Cook's distance greater than 1).

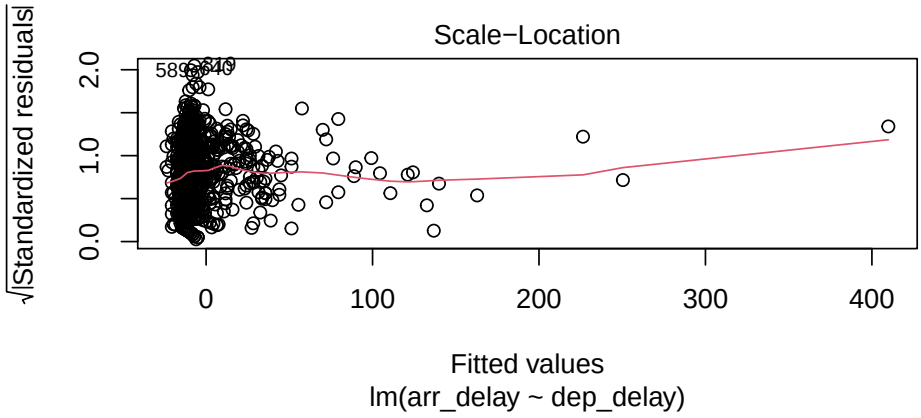
```
delay_model %>%
  plot(which = c(1))
```



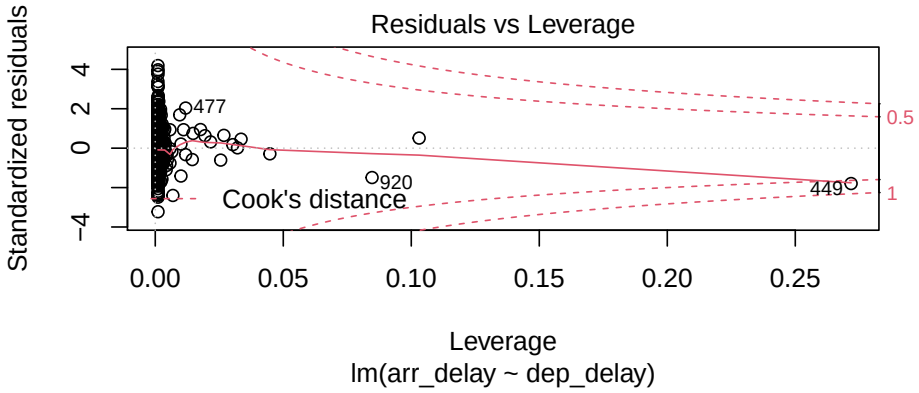
```
delay_model %>%
  plot(which = c(2))
```




```
delay_model %>%
  plot(which = c(3))
```




```
delay_model %>%
  plot(which = c(5))
```



8.1.3 How to report

Overall, we can say that the delay model computed above is fit ($F(1, 972) = 3396.74$, $p < .001$), indicating that the departure delay might account for 77.73% of the arrival delay. However the model is only partially robust. The residuals satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 0.02$, $p = 0.9$), and the independence assumption (Durbin-Watson test, $DW = 1.87$, $p = 0.02$), but they are not normally distributed (Shapiro-Wilk test, $W = 0.98$, $p < .001$).

The `stargazer` function of the `stargazer` library can be applied to the model `delay_model` to generate a nicer output in RMarkdown PDF documents by including `results = "asis"` in the R snippet option.

```
# Install stargazer if not yet installed
# install.packages("stargazer")

library(stargazer)

# Not rendered in bookdown
stargazer(delay_model, header = FALSE)
```

8.2 Multiple regression

The multiple regression analysis is a supervised machine learning approach to creating a model able to predict the value of one outcome variable Y based on two or more predictor variables $X_1 \dots X_M$, by estimating the intercept b_0 and the coefficients (slopes) $b_1 \dots b_M$, and accounting for a reasonable amount of error ϵ.

$$Y_i = (b_0 + b_1 * X_{i1} + b_2 * X_{i2} + \cdots + b_M * X_{iM}) + \epsilon_i$$

Table 8.1:

	Dependent variable:
	arr_delay
dep_delay	1.042***
	(0.018)
Constant	−4.967***
	(0.437)
Observations	974
R^2	0.778
Adjusted R^2	0.777
Residual Std. Error	13.618 (df = 972)
F Statistic	3,396.742*** (df = 1; 972)
Note:	*p<0.1; **p<0.05; ***p<0.01

The assumptions are the same as the simple regression, plus the assumption of **no multicollinearity**: if two or more predictor variables are used in the model, each pair of variables not correlated. This assumption can be tested by checking the variance inflation factor (VIF). If the largest VIF value is greater than 10 or the average VIF is substantially greater than 1, there might be an issue of multicollinearity.

8.2.1 Example

The example below explores whether a regression model can be created to estimate the number of people in Leicester commuting to work using private transport (`u121`) in Leicester, using the number of people in different industry sectors as predictors.

For instance, occupations such as electricity, gas, steam and air conditioning supply (`u144`) require to travel some distances with equipment, thus the related variable `u144` is included in the model, whereas people working in information and communication might be more likely to work from home or commute by public transport.

A multiple regression model can be specified in a similar way as a simple regression model, using the same `lm` function, but adding the additional predictor variables using a `+` operator.

```
leicester_2011OAC <- readr::read_csv("2011_OAC_Raw_uVariables_Leicester.csv")
```

```
# Select and
# normalise variables
```

```r
leicester_2011OAC_transp <-
  leicester_2011OAC %>%
  dplyr::select(
    OA11CD,
    Total_Pop_No_NI_Students_16_to_74, Total_Employment_16_to_74,
    u121, u141:u158
  ) %>%
  # percentage method of travel
  dplyr::mutate(
    u121 = (u121 / Total_Pop_No_NI_Students_16_to_74) * 100
  ) %>%
  # percentage across industry sector columns
  dplyr::mutate(
    dplyr::across(
      u141:u158,
      function(x){ (x / Total_Employment_16_to_74) * 100 }
    )
  ) %>%
  # rename columns
  dplyr::rename_with(
    function(x){ paste0("perc_", x) },
    c(u121, u141:u158)
  )

# Selected variables

# perc_u120: Method of Travel to Work, Private Transport
# perc_u142: Industry Sector, Mining and quarrying
# perc_u144: Industry Sector, Electricity, gas, steam and air conditioning ...
# perc_u146: Industry Sector, Construction
# perc_u149: Industry Sector, Accommodation and food service activities

# Create model
commuting_model1 <-
  leicester_2011OAC_transp %$%
  lm(
    perc_u121 ~
      perc_u142 + perc_u144 + perc_u146 + perc_u149
  )

# Print summary
commuting_model1 %>%
  summary()
```

```
##
```

```
## Call:
## lm(formula = perc_u121 ~ perc_u142 + perc_u144 + perc_u146 +
##     perc_u149)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.315  -6.598  -0.244   6.439  31.472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.12690    0.94148  39.434  < 2e-16 ***
## perc_u142    3.74768    1.21255   3.091  0.00205 **
## perc_u144    1.16865    0.25328   4.614 4.48e-06 ***
## perc_u146    1.05408    0.09335  11.291  < 2e-16 ***
## perc_u149   -1.56948    0.08435 -18.606  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.481 on 964 degrees of freedom
## Multiple R-squared:  0.3846,Adjusted R-squared:  0.3821
## F-statistic: 150.6 on 4 and 964 DF,  p-value: < 2.2e-16
```

```
# Not rendered in bookdown
stargazer(commuting_model1, header=FALSE)
```

```
commuting_model1 %>%
  rstandard() %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99889, p-value = 0.8307
```

```
commuting_model1 %>%
  bptest()
```

```
##
##  studentized Breusch-Pagan test
##
## data:  .
## BP = 28.403, df = 4, p-value = 1.033e-05
```

```
commuting_model1 %>%
  dwtest()
```

```
##
```

Table 8.2:

	Dependent variable:
	perc_u121
perc_u142	3.748***
	(1.213)
perc_u144	1.169***
	(0.253)
perc_u146	1.054***
	(0.093)
perc_u149	−1.569***
	(0.084)
Constant	37.127***
	(0.941)
Observations	969
R^2	0.385
Adjusted R^2	0.382
Residual Std. Error	9.481 (df = 964)
F Statistic	150.622*** (df = 4; 964)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##  Durbin-Watson test
##
## data:  .
## DW = 1.835, p-value = 0.004908
## alternative hypothesis: true autocorrelation is greater than 0
```

```
library(car)

commuting_model1 %>%
  vif()
```

```
## perc_u142 perc_u144 perc_u146 perc_u149
##  1.006906  1.016578  1.037422  1.035663
```

The output above suggests that the model is fit ($F(4, 964) = 150.62$, $p < .001$), indicating that a model based on the presence of people working in the four selected industry sectors can account for 38.21% of the number of people using private transportation to commute to work. However the model is only partially robust. The residuals are normally distributed (Shapiro-Wilk test, $W = 1$, $p = 0.83$) and there seems to be no multicollinearity with average VIF 1.02, but the residuals don't satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 28.4$, $p < .001$), nor the independence assumption (Durbin-Watson test, $DW = 1.84$, $p < .01$).

The coefficient values calculated by the `lm` functions are important to create the model, and provide useful information. For instance, the coefficient for the variable `perc_u144` is `1.169`, which indicates that if the presence of people working in electricity, gas, steam and air conditioning supply increases by one percentage point, the number of people using private transportation to commute to work increases by `1.169` percentage points, according to the model. The coefficients also indicate that the presence of people working in accommodation and food service activities actually has a negative impact (in the context of the variables selected for the model) on the number of people using private transportation to commute to work.

In this example, all variables use the same unit and are of a similar type, which makes interpreting the model relatively simple. When that is not the case, it can be useful to look at the standardized β, which provide the same information but measured in terms of standard deviation, which make comparisons between variables of different types easier to draw. For instance, the values calculated below using the function `lm.beta` of the library `lm.beta` indicate that if the presence of people working in construction has the highest impact on the outcome varaible. If the presence of people working in construction increases by one standard deviation, the number of people using private transportation to commute to work increases by `0.29` standard deviations, according to the model.

```
# Install lm.beta library if necessary
# install.packages("lm.beta")
```

```
library(lm.beta)

lm.beta(commuting_model1)
```

```
##
## Call:
## lm(formula = perc_u121 ~ perc_u142 + perc_u144 + perc_u146 +
##     perc_u149)
##
## Standardized Coefficients::
## (Intercept)   perc_u142   perc_u144   perc_u146   perc_u149
##  0.00000000  0.07836017  0.11754058  0.29057993 -0.47841083
```

8.3 Exercise 324.1

Question 324.1.1: Create a model having as outcome variable the presence of people using private transport for commuting to work, and using a stepwise *"both"* approach, having all the variables created for the example above and related to the presence of people working in different industry sectors (`perc_u141 to perc_u158`) as scope.

Question 324.1.2: Is the presence of people using public transportation to commute to work statistically, linearly related to mean age (`u020`)?

Question 324.1.3: Is the presence of people using public transportation to commute to work statistically, linearly related to (a subset of) the age structure categories (`u007` to `u019`)?

Chapter 9

Supervised machine learning

Stefano De Sabbata

9.1 Introduction

The field of **machine learning** sits at the intersection of computer science and statistics, and it is a core component of data science. According to Mitchell (1997), *"the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."*

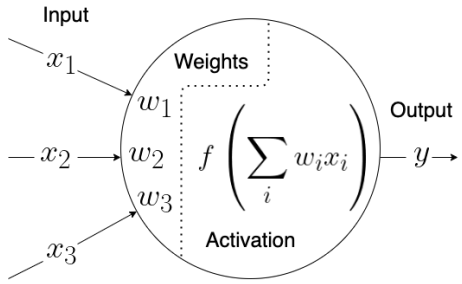
Machine learning approaches are divided into two main types.

- **Supervised**:
  - training of a *"predictive"* model from data;
  - one (or more) attribute of the dataset is used to "predict" another attribute.
- **Unsupervised**:
  - discovery of *descriptive* patterns in data;
  - commonly used in data mining.

Classification is one of the classic supervised machine learning tasks, where algorithms are used to learn (i.e., model) the relationship between a series of input values (a.k.a. predictors, independent variables) and output categorical values or labels (a.k.a. outcome, dependent variable). A model trained on a training dataset can learn the relationship between the input and the labels, and then be used to label new, unlabeled data.

9.1.1 Artificial neural networks

Artificial neural networks (ANNs) are one of the most studied approaches in supervised machine learning, and the term actually defines a large set of different approaches. These model aims to simulate a simplistic version of a brain made of artificial neurons. Each artificial neuron combines a series of input values into one output values, using a series of weights (one per input value) and an activation function. The aim of the model is to learn an optimal set of weights that, once combined with the input values, generates the correct output value. The latter is also influenced by the activation function, which modulates the final result.



Each neuron is effectively a regression model. The input values are the predictors (or independent variables), the output is the outcome (or dependent variable), and the weights are the coefficients (see also previous practical on regression models). The selection of the activation function defines the regression model. As ANNs are commonly used for classification, one of the most common activation functions used is the sigmoid, thus rendering every single neuron a logistic regression.

An instance of an ANN is defined by its topology (number of layers and nodes), activation functions and the algorithm used to train the network. The selection of all those parameters renders the construction of ANNs a very complex task, and the quality of the result frequently relies on the experience of the data scientist.

- Number of layers
  - Single-layer network: one node of input variable one node per category of the output variable, effectively a logistic regression.
  - Multi-layer network: adds one hidden layer, which aims to capture hidden *"features"* of the data, as combinations of the input values, and use that for the final classification.
  - Deep neural networks: several hidden layers, each aiming to capture more and more complex *"features"* of the data.
- Number of nodes
  - The number of nodes needs to be selected for each one of the hidden layers.

9.1.2 Support vector machines

Support vector machines (SVMs) are another very common approach to supervised classification. SVMs perform the classification task by partitioning the data space into regions separated by hyperplanes. For instance, in a bi-dimensional space, a hyperplane is a line, and the algorithm is designed to find the line that best separates two groups of data. Computationally, the process is not dissimilar to linear regression.

9.1.3 Confusion matrices

Once a classification model has been created, the next step is validation. The latter can involve different approaches and procedures, but one of the most common and simple approaches is to split the data between a training and a testing set. The model is trained on the training set and then validated using the testing set. Both sets will contain both the input values (predictors) and the output values (outcome).

The model trained using the training set can be used to predict the values for the testing set. The outcome of the prediction can be compared to the actual categories in the testing dataset. A **confusion matrix** is a representation of the correspondence between actual values and predicted values in the testing dataset, including:

- true positive: correctly classified as the first (positive) class;
- true negative: correctly classified as the second (negative) class;
- false positive: incorrectly classified as the first (positive) class;
- false negative: incorrectly classified as the second (negative) class.

The number of true and false positive and negatives are used to calculate a number of performance measures. The simplest measures of performance are *accuracy* and *error rate*.

$$accuracy = \frac{true\ positive + true\ negative}{total\ number\ of\ cases}$$

$$error\ rate = 1 - accuracy$$

There are also a number of additional measures that can provide further insight into the quality of the prediction, such as *sensitivity* (true positive rate) and *specificity* (true negative rate). If the model has been created to predict a binary categorical variable, based on the definition above (the first category is positive, the second category is negative), sensitivity is a measure of quality in predicting the first category, and specificity is a measure of quality in predicting the second category.

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} = \frac{correct\ 1st}{all\ 1st}$$

$$specificity = \frac{true\ negative}{true\ negative + false\ positive} = \frac{correct\ 2nd}{all\ 2nd}$$

Two further, similar measures are *precision* and *recall*. A model with high precision is a model that can be trusted to make a correct prediction when identifying an observation as being part of the first category. The formula for recall is the same used for sensitivity, but in this case, it has a different interpretation, derived from the computer science literature on search engines, where a model with high recall is able to correctly retrieve most items of the specified category. Note that both precision and recall are dependent on which one of two categories is defined as being the first.

$$precision = \frac{true\ positive}{true\ positive + false\ positive} = \frac{correct\ 1st}{predicted\ as\ 1st}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} = \frac{correct\ 1st}{all\ 1st}$$

Precision and recall can also be combined into a single measure of performance called *F-score* (a.k.a., *F-measure* or *F1*).

$$F - score = \frac{2 \times precision \times recall}{precision + recall}$$

Finally, the *kappa statistic* (the most common being Cohen's kappa) is an additional measure of accuracy, which measures the agreement between prediction and actual values, while also accounting for the probability of correct prediction by chance.

9.2 Examples

The two examples below explore the relation between some of the variables from the United Kingdom 2011 Census included among the 167 initial variables used to create the 2011 Output Area Classification (Gale *et al.*, 2016) and the Rural Urban Classification (2011) of Output Areas in England and Wales created by the Office for National Statistics. The various examples and models explore whether it is possible to learn the rural-urban distinction by using some of those census variables, in the Local Authority Districts (LADs) in Leicestershire (excluding the city of Leicester itself.

The code below uses the libraries `caret`, `e1071` and `neuralnet`. Please install them before continuing with the practical.

```
install.packages("caret")
install.packages("e1071")
install.packages("neuralnet")
```

9.2.1 Data

The examples use the same data seen in previous practicals, but for the 7 LADs in Leicestershire outside the boundaries of the city of Leicester: Blaby, Charnwood, Harborough, Hinckley and Bosworth, Melton, North West Leicestershire, and Oadby and Wigston. Those data are loaded from the `2011_OAC_Raw_uVariables_Leicestershire.csv`. The second part of the code extracts the data of the Rural Urban Classification (2011) from the compressed file `RUC11_OA11_EW.zip`, loads the extracted data and finally deletes them.

```
# Libraries
library(tidyverse)
library(magrittr)

# 2011 OAC data for Leicestershire (excl. Leicester)
liec_shire_2011OAC <- readr::read_csv("2011_OAC_Raw_uVariables_Leicestershire.csv")

# Rural Urban Classification (2011)
# >>> Note that if you upload the file to RStudio Server
# >>> the file will be automatically unzipped
# >>> thus the unzip and unlink instrcutions are not necessary
unzip("RUC11_OA11_EW.zip")
ru_class_2011 <- readr::read_csv("RUC11_OA11_EW.csv")
unlink("RUC11_OA11_EW.csv")
```

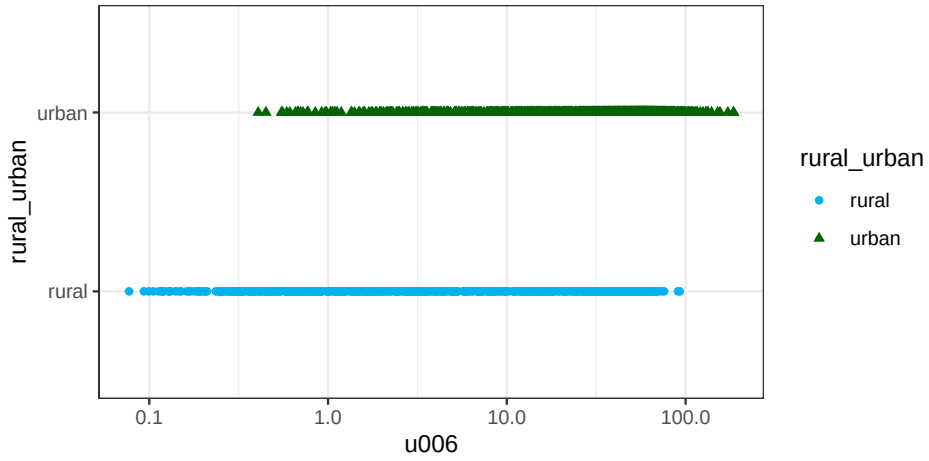
We can then join the two datasets and create a simplified, binary rural - urban classification, that is used in the examples below.

```
liec_shire_2011OAC_RU <-
  liec_shire_2011OAC %>%
  dplyr::left_join(ru_class_2011) %>%
  dplyr::mutate(
    rural_urban =
      forcats::fct_recode(
        RUC11CD,
        urban = "C1",
        rural = "D1",
        rural = "E1",
        rural = "F1"
      ) %>%
      forcats::fct_relevel(
        c("rural", "urban")
      )
  )
```

9.2.2 Logistic regression

Can we predict whether an Output Area (OA) is urban or rural, solely based on its population density?

```
liec_shire_2011OAC_RU %>%
  ggplot2::ggplot(
    aes(
      x = u006,
      y = rural_urban
    )
  ) +
  ggplot2::geom_point(
    aes(
      color = rural_urban,
      shape = rural_urban
    )
  ) +
  ggplot2::scale_color_manual(values = c("deepskyblue2", "darkgreen")) +
  ggplot2::scale_x_log10() +
  ggplot2::theme_bw()
```



The two patters in the plot above seem quite close even when plotted using a logarithmically transformed x-axis. As a first step, we can extract from the dataset only the data we need, and create a logarithmic transformation of the population density value. To be able to perform a simple validatin our model, we can divide that data in a training (80% of the dataset) and a testing set (20% of the dataset).

```
# Data for logit model
ru_logit_data <-
  liec_shire_2011OAC_RU %>%
```

```
  dplyr::select(OA11CD, u006, rural_urban) %>%
  dplyr::mutate(
    density_log = log10(u006)
  )

# Training set
ru_logit_data_trainig <-
  ru_logit_data %>%
  slice_sample(prop = 0.8)

# Testing set
ru_logit_data_testing <-
  ru_logit_data %>%
  anti_join(ru_logit_data_trainig)
```

We can then compute the logit model using the `stats::glm` function and specifying `binomial()` as `family`. The summary of the model highlights how the model is significant, but `Residual deviance` is fairly close to the `Null deviance` (null model), which is not a good sign.

```
ru_logit_model <-
  ru_logit_data_trainig %$%
  stats::glm(
    rural_urban ~
      density_log,
    family = binomial()
  )

ru_logit_model %>%
  summary()
```

```
##
## Call:
## stats::glm(formula = rural_urban ~ density_log, family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1830  -0.5822   0.5335   0.6586   2.0383
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3082     0.1322  -9.896   <2e-16 ***
## density_log   1.8273     0.1008  18.128   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2063.6  on 1667  degrees of freedom
## Residual deviance: 1601.7  on 1666  degrees of freedom
## AIC: 1605.7
##
## Number of Fisher Scoring iterations: 4
```

As per other regression models, it would be necessary to test the assumptions of the logit model and the overall distribution of the residuals. However, as this model only has one predictor, we will confine our performance analysis to simple validation.

We can test the performance of the model through a validation exercise, using the testing dataset. Finally, we can compare the results of the prediction with the original data using a confusion matrix.

```
ru_logit_prediction <-
  ru_logit_model %>%
  # Use model to predict values
  stats::predict(
    ru_logit_data_testing,
    type = "response"
  ) %>%
  as.numeric()

ru_logit_data_testing <-
  ru_logit_data_testing %>%
  tibble::add_column(
    # Add column with predicted class
    logit_predicted_ru =
      # Values below 0.5 indicate first factor level (rural)
      # Values above 0.5 indicate second factor level (ruban)
      ifelse(
        ru_logit_prediction <= 0.5,
        "rural", # first factor level
        "urban"  # second factor level
      ) %>%
      forcats::as_factor() %>%
      forcats::fct_relevel(
        c("rural", "urban")
      )
  )

# Load library for confusion matrix
library(caret)
```

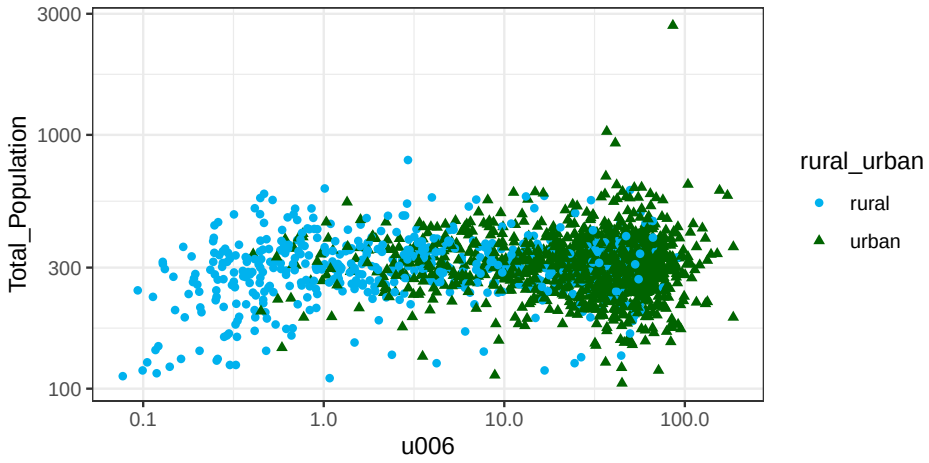
```
# Confusion matrix
caret::confusionMatrix(
  ru_logit_data_testing %>% dplyr::pull(logit_predicted_ru),
  ru_logit_data_testing %>% dplyr::pull(rural_urban),
  mode = "everything"
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction rural urban
##      rural    72    28
##      urban    58   259
##
##                Accuracy : 0.7938
##                  95% CI : (0.7517, 0.8316)
##     No Information Rate : 0.6882
##     P-Value [Acc > NIR] : 9.325e-07
##
##                   Kappa : 0.487
##
##  Mcnemar's Test P-Value : 0.001765
##
##             Sensitivity : 0.5538
##             Specificity : 0.9024
##          Pos Pred Value : 0.7200
##          Neg Pred Value : 0.8170
##               Precision : 0.7200
##                  Recall : 0.5538
##                      F1 : 0.6261
##              Prevalence : 0.3118
##          Detection Rate : 0.1727
##    Detection Prevalence : 0.2398
##       Balanced Accuracy : 0.7281
##
##        'Positive' Class : rural
##
```

9.2.3 Support vector machines

To showcase the use of SVMs, the example below expands on the one above by building a model for urban - rural classification that uses total population and area (logarithmically transformed) as two separate input values, rather than combined as population density. The aim of the SVM is then to find a line that maximises the margin between the two groups shown in the plot below.

```
liec_shire_2011OAC_RU %>%
  ggplot2::ggplot(
    aes(
      x = u006,
      y = Total_Population
    )
  ) +
  ggplot2::geom_point(
    aes(
      color = rural_urban,
      shape = rural_urban
    )
  ) +
  ggplot2::scale_color_manual(values = c("deepskyblue2", "darkgreen")) +
  ggplot2::scale_x_log10() +
  ggplot2::scale_y_log10() +
  ggplot2::theme_bw()
```



The plot illustrates how the two variables are skewed (note that the axes are logarithmically transformed) and that the two groups are not linearly separable. We can thus follow a procedure similar to the one seen above: extract the necessary data; split the data between training and testing for validation; build the model; predict the values for the testing set and interpret the confusion matrix.

```
# Data for SVM model
ru_svm_data <-
  liec_shire_2011OAC_RU %>%
  dplyr::select(OA11CD, Total_Population, u006, rural_urban) %>%
  dplyr::mutate(
    area_log = log10(u006),
```

```
    population_log = log10(Total_Population)
  )

# Training set
ru_svm_data_trainig <-
  ru_svm_data %>%
  slice_sample(prop = 0.8)

# Testing set
ru_svm_data_testing <-
  ru_svm_data %>%
  anti_join(ru_svm_data_trainig)

# Load library for svm function
library(e1071)

# Build the model
ru_svm_model <-
  ru_svm_data_trainig %$%
  e1071::svm(
    rural_urban ~
      area_log + population_log,
    # Use a simple linear hyperplane
    kernel = "linear",
    # Scale the data
    scale = TRUE,
    # Cost value for observations
    # crossing the hyperplane
    cost = 10
  )

# Predict the values for the testing dataset
ru_svm_prediction <-
  stats::predict(
    ru_svm_model,
    ru_svm_data_testing %>%
      dplyr::select(area_log, population_log)
  )

# Add predicted values to the table
ru_svm_data_testing <-
  ru_svm_data_testing %>%
  tibble::add_column(
    svm_predicted_ru = ru_svm_prediction
  )
```

```
# Confusion matrix
caret::confusionMatrix(
  ru_svm_data_testing %>% dplyr::pull(svm_predicted_ru),
  ru_svm_data_testing %>% dplyr::pull(rural_urban),
  mode = "everything"
  )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction rural urban
##      rural    67    18
##      urban    69   263
##
##                Accuracy : 0.7914
##                  95% CI : (0.7492, 0.8294)
##     No Information Rate : 0.6739
##     P-Value [Acc > NIR] : 7.190e-08
##
##                   Kappa : 0.4745
##
##  Mcnemar's Test P-Value : 8.296e-08
##
##             Sensitivity : 0.4926
##             Specificity : 0.9359
##          Pos Pred Value : 0.7882
##          Neg Pred Value : 0.7922
##               Precision : 0.7882
##                  Recall : 0.4926
##                      F1 : 0.6063
##              Prevalence : 0.3261
##          Detection Rate : 0.1607
##    Detection Prevalence : 0.2038
##       Balanced Accuracy : 0.7143
##
##        'Positive' Class : rural
##
```

As a third more complex example, we can explore how to build a model for rural - urban classification using the presence of the five different types of dwelling as input variables. The confusion matrix below clearly illustrates that a simple linear SVM is not adequate to construct such a model.

```
# Data for SVM model
ru_dwellings_data <-
  liec_shire_2011OAC_RU %>%
```

```
  dplyr::select(
    OA11CD, rural_urban, Total_Dwellings,
    u086:u090
  ) %>%
  # scale across
  dplyr::mutate(
    dplyr::across(
      u086:u090,
      scale
      #function(x){ (x / Total_Dwellings) * 100 }
    )
  ) %>%
  dplyr::rename(
    scaled_detached = u086,
    scaled_semidetached = u087,
    scaled_terraced = u088,
    scaled_flats = u089,
    scaled_carava_tmp = u090
  ) %>%
  dplyr::select(-Total_Dwellings)

# Training set
ru_dwellings_data_trainig <-
  ru_dwellings_data %>%
  slice_sample(prop = 0.8)

# Testing set
ru_dwellings_data_testing <-
  ru_dwellings_data %>%
  anti_join(ru_dwellings_data_trainig)

# Build the model
ru_dwellings_svm_model <-
  ru_dwellings_data_trainig %$%
  e1071::svm(
    rural_urban ~
      scaled_detached + scaled_semidetached + scaled_terraced +
      scaled_flats + scaled_carava_tmp,
    # Use a simple linear hyperplane
    kernel = "linear",
    # Cost value for observations
    # crossing the hyperplane
    cost = 10
  )
```

```r
# Predict the values for the testing dataset
ru_dwellings_svm_prediction <-
  stats::predict(
    ru_dwellings_svm_model,
    ru_dwellings_data_testing %>%
      dplyr::select(scaled_detached:scaled_carava_tmp)
  )

# Add predicted values to the table
ru_dwellings_data_testing <-
  ru_dwellings_data_testing %>%
  tibble::add_column(
    dwellings_svm_predicted_ru = ru_dwellings_svm_prediction
  )

# Confusion matrix
caret::confusionMatrix(
  ru_dwellings_data_testing %>% dplyr::pull(dwellings_svm_predicted_ru),
  ru_dwellings_data_testing %>% dplyr::pull(rural_urban),
  mode = "everything"
  )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction rural urban
##      rural     0     0
##      urban   126   291
##
##                Accuracy : 0.6978
##                  95% CI : (0.6513, 0.7416)
##     No Information Rate : 0.6978
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.0000
##             Specificity : 1.0000
##          Pos Pred Value :    NaN
##          Neg Pred Value : 0.6978
##               Precision :     NA
##                  Recall : 0.0000
##                      F1 :     NA
```

```
##              Prevalence : 0.3022
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : rural
##
```

9.2.4 Kernels

Instead of relying on simple linear hyperplanes, we can use the *"kernel trick"* to project the data into a higher-dimensional space, which might allow groups to be (more easily) linearly separable.

```
# Build a second model
# using a radial kernel
ru_dwellings_svm_radial_model <-
  ru_dwellings_data_trainig %$%
  e1071::svm(
    rural_urban ~
      scaled_detached + scaled_semidetached + scaled_terraced +
      scaled_flats + scaled_carava_tmp,
    # Use a radial kernel
    kernel = "radial",
    # Cost value for observations
    # crossing the hyperplane
    cost = 10
  )

# Predict the values for the testing dataset
ru_svm_dwellings_radial_prediction <-
  stats::predict(
    ru_dwellings_svm_radial_model,
    ru_dwellings_data_testing %>%
      dplyr::select(scaled_detached:scaled_carava_tmp)
  )

# Add predicted values to the table
ru_dwellings_data_testing <-
  ru_dwellings_data_testing %>%
  tibble::add_column(
    dwellings_radial_predicted_ru = ru_svm_dwellings_radial_prediction
  )

# Confusion matrix
caret::confusionMatrix(
```

```
  ru_dwellings_data_testing %>% dplyr::pull(dwellings_radial_predicted_ru),
  ru_dwellings_data_testing %>% dplyr::pull(rural_urban),
  mode = "everything"
  )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction rural urban
##      rural    41    29
##      urban    85   262
##
##                Accuracy : 0.7266
##                  95% CI : (0.6811, 0.7689)
##     No Information Rate : 0.6978
##     P-Value [Acc > NIR] : 0.1093
##
##                   Kappa : 0.2583
##
##  Mcnemar's Test P-Value : 2.588e-07
##
##             Sensitivity : 0.32540
##             Specificity : 0.90034
##          Pos Pred Value : 0.58571
##          Neg Pred Value : 0.75504
##               Precision : 0.58571
##                  Recall : 0.32540
##                      F1 : 0.41837
##              Prevalence : 0.30216
##          Detection Rate : 0.09832
##    Detection Prevalence : 0.16787
##       Balanced Accuracy : 0.61287
##
##        'Positive' Class : rural
##
```

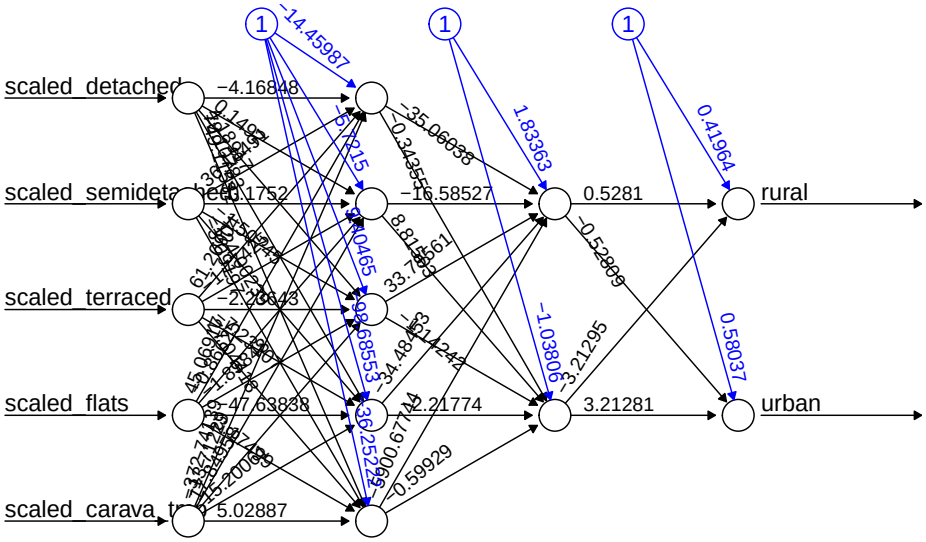
9.2.5 Artificial neural network

Finally, we can explore the construction of an ANN for the same input and output variables seen in the previous example, and compare which one of the two approaches produces better results. The example below creates a multi-layer ANN, using two hidden layers, with five and two nodes, respectively.

```
# Load library for ANNs
library(neuralnet)
```

```
# Build a third model
# using an ANN
ru_dwellings_nnet_model <-
  neuralnet::neuralnet(
    rural_urban ~
      scaled_detached + scaled_semidetached + scaled_terraced +
      scaled_flats + scaled_carava_tmp,
    data = ru_dwellings_data_trainig,
    # Use 2 hidden layers
    hidden = c(5, 2),
    # Max num of steps for training
    stepmax = 1000000
  )
```

```
ru_dwellings_nnet_model %>%  plot(rep = "best")
```



Error: 278.244106 Steps: 60553

```
# Predict the values for the testing dataset
ru_dwellings_nnet_prediction <-
  neuralnet::compute(
    ru_dwellings_nnet_model,
    ru_dwellings_data_testing %>%
      dplyr::select(scaled_detached:scaled_carava_tmp)
  )

# Derive predicted categories
ru_dwellings_nnet_predicted_categories <-
  # from the prediction object
```

```
  ru_dwellings_nnet_prediction %$%
  # extract the result
  # which is a matrix of probabilities
  # for each object and category
  net.result %>%
  # select the column (thus the category)
  # with higher probability
  max.col %>%
  # recode columns values as
  # rural or urban
  dplyr::recode(
    `1` = "rural",
    `2` = "urban"
  ) %>%
  forcats::as_factor() %>%
  forcats::fct_relevel(
    c("rural", "urban")
  )

# Add predicted values to the table
ru_dwellings_data_testing <-
  ru_dwellings_data_testing %>%
  tibble::add_column(
    dwellings_nnet_predicted_ru =
      ru_dwellings_nnet_predicted_categories
  )

# Confusion matrix
caret::confusionMatrix(
  ru_dwellings_data_testing %>% dplyr::pull(dwellings_nnet_predicted_ru),
  ru_dwellings_data_testing %>% dplyr::pull(rural_urban),
  mode = "everything"
  )
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction rural urban
##      rural    46    34
##      urban    80   257
##
##                Accuracy : 0.7266
##                  95% CI : (0.6811, 0.7689)
##     No Information Rate : 0.6978
##     P-Value [Acc > NIR] : 0.1093
```

```
##
##                   Kappa : 0.2769
##
##  Mcnemar's Test P-Value : 2.502e-05
##
##             Sensitivity : 0.3651
##             Specificity : 0.8832
##          Pos Pred Value : 0.5750
##          Neg Pred Value : 0.7626
##               Precision : 0.5750
##                  Recall : 0.3651
##                      F1 : 0.4466
##              Prevalence : 0.3022
##          Detection Rate : 0.1103
##    Detection Prevalence : 0.1918
##       Balanced Accuracy : 0.6241
##
##        'Positive' Class : rural
##
```

9.3 Exercise 404.1

Question 404.1.1: Create an SVM model capable of classifying areas in Leicester and Leicestershire as rural or urban based on the series of variables that relate to *"Economic Activity"* among the 167 initial variables used to create the 2011 Output Area Classification (Gale *et al.*, 2016).

Question 404.1.2: Create an ANN using the same input and output values used in Question 404.1.1.

Question 404.1.3: Assess which one of the two models preforms a better classification.

Chapter 10

Unsupervised machine learning

Stefano De Sabbata

10.1 Introduction

The field of **machine learning** sits at the intersection of computer science and statistics, and it is a core component of data science. According to Mitchell (1997), *"the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."*

Machine learning approaches are divided into two main types.

- **Supervised**:
  - training of a *"predictive"* model from data;
  - one (or more) attribute of the dataset is used to "predict" another attribute.
- **Unsupervised**:
  - discovery of *descriptive* patterns in data;
  - commonly used in data mining.

Clustering is a classic unsupervised machine learning task, which aims to *"automatically divides the data into* ***clusters*** *, or groups of similar items"*(Lantz, 2019). In computer science, a wide range of approaches has been developed to tackle clustering. Among those approaches, the two most common are centroid-based approaches (such as k-means) and hierarchical approaches. Other approaches include density-based clustering methods (such as DBSCAN) and

mixed approaches (such as bagged clustering), which combine different aspects of centroid-based and hierarchical approaches.

10.1.1 K-means

The k-mean approach clusters n observations (x) in k clusters (c) by minimising the within-cluster sum of squares (WCSS) through an iterative process. That is, the algorithm calculates the distance between each observation (i.e., each case, object, row in the table) and the centroid of its cluster. The square values of those distances are summed up for each cluster, and then for the whole dataset. The aim of the algorithm is minimise that value.

$$WCSS = \sum_{c=1}^{k} \sum_{x \in c} (x - \overline{x}_c)^2$$

To minimise WCSS, while trying to identify `k` clusters, k-mean first randomly select `k` observations as initial centroids. Then, k-means repeats the two steps below. Every time k-means repeats those two steps, the new centroids will be closer to the two actual centre. The process continues until centroids don't change anymore (within a certain margin of error) or until it has reached a maximum number of iterations set by the analyst.

- **assignment step**: observations assigned to closest centroids
- **update step**: calculate means for each cluster, as new the centroid

10.1.2 Number of clusters

A key limitation of k-mean is that it requires to select the number of clusters to be identified in advance. Unfortunately, analysts are not always in the position of knowing in advance how many clusters are there supposed to be within the data they are analysing. In such cases, there are a number of heuristics that can be used to select the number of clusters that best fits the data.

The most well-known method is the *"elbow method"*. This approach suggests to calculate k-means for a range of values of k, calculate the WCSS obtained for each k, and then select the value of k that minimises WCSS without increasing the number of clusters beyond the point where the decrease in WCSS is minimal. This approach is called *"elbow method"* because (as can be seen in the examples below) when printing a line representing the value of WCSS for all values of k taken into account, it suggests to select the value of k at the *"elbow"* or inflation point of the line.

Other heuristics exist, which suggest using alternative measures of cluster quality. For instance, the `cluster` library provides simple ways to calculate the silhouette measure and the gap statistic. The silhouette value indicates how well observations fit within their clusters, whereas the gap statistic measures

the dispersion within each cluster, compared to a uniform distribution of values. The higher the value of the gap statistic, the further away the distribution is from uniform (thus the higher the quality of the clustering).

For all three heuristics, the best approach would be to calculate those values using a bootstrapping approach. That is, to calculate the same statistics multiple times on samples of the dataset, in order to account for random variation. However, only the `clusGap` function of the `cluster` library allows for bootstrapping natively, as illustrated below.

```
library(cluster)
```

10.1.3 Geodemographic Classification

In GIScience, clustering approaches are commonly used to create *geodemographic classifications*. For instance, Gale *et al.*, 2016 created the 2011 Output Area Classification (2011 OAC) starting from an initial set of 167 prospective variables from the UK Census 2011.

In the process of creating the classification, 86 variables were removed from the initial set, including highly correlated variables that don't bring additional information to the classification process. Furthermore, 41 variable were retained as they were, whereas 40 were combined, to create final set of 60 variables. The k-mean approach was then applied to cluster the census Output Areas (OAs) into 8 supergroups, 26 groups and 76 subgroups.

The paper provides a detail report of the process. In particular, it is interesting to see how the authors applied a process of variable selection involving repeated clustering while excluding one variable, to see how the within cluster sum of square measure (WCSS) would be affected. Variable that produced significantly higher WCSS when excluded were considered for exclusion from the final analysis, in order to increase the homogeneity of the clusters.

Once the clustering is completed, the final step in geodemographic classification is the interpretation of the resulting cluster, which is commonly done by observing the average values of the variables for each cluster.

10.2 Examples

The two examples below explore the creation of simple geodemographic classifications for the city of Leicester, using a few variables from the United Kingdom 2011 Census and included among the 167 initial variables that Gale *et al.*, 2016 have taken into account in creating the 2011 Output Area Classification.

```
leicester_2011OAC <- readr::read_csv("2011_OAC_Raw_uVariables_Leicester.csv")
```

The variables that we are going to take into account are the five ones listed below, plus the total count for their statistical unit, that is `Total_Dwellings`.

- u086: Detached
- u087: Semi-detached
- u088: Terraced (including end-terrace)
- u089: Flats
- u090: Caravan or other mobile or temporary structure

The code below extracts the necessary variables from the original dataset and applies the normalisation steps across all the five variables listed above. Finally, the columns are renamed with more user-friendly names and adding `perc_` in front of the column name.

```
leicester_dwellings <-
  leicester_2011OAC %>%
  dplyr::select(
    OA11CD, Total_Dwellings,
    u086:u090
  ) %>%
  # scale across
  dplyr::mutate(
    dplyr::across(
      u086:u090,
      #scale
      function(x){ (x / Total_Dwellings) * 100 }
    )
  ) %>%
  dplyr::rename(
    detached = u086,
    semidetached = u087,
    terraced = u088,
    flats = u089,
    carava_tmp = u090
  ) %>%
  # rename columns
  dplyr::rename_with(
    function(x){ paste0("perc_", x) },
    detached:carava_tmp
  )
```

The first step of k-means is to select some observations at random as initial centroids. As a result, every time we run the computation, the starting point will be slightly different, and so might be the result. In particular, it is likely that the cluster order might chance (i.e., what was cluster `1` one time might be cluster `3` the next), although the overall result should be stable. Nevertheless, to make this document more reproducible, we can set a *"seed"* for the generation of random numbers. That will ensure that the observations selected at random will be the same, and thus the results will be the same. That can be done in R, using the function `set.seed` and providing a number (any relatively large

number will do) as *"seed"*.

```
set.seed(20201208)
```

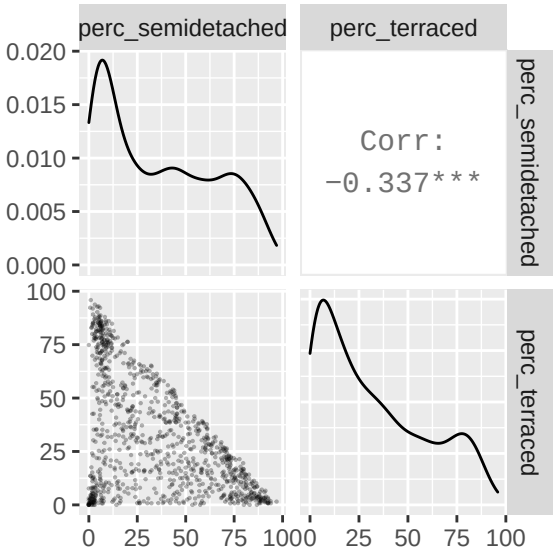
10.2.1 Terrace and semi-detached houses

The first example explores how to create a geodemographic classification using only two variables.

- `u087` (now `semidetached`): Semi-detached
- `u088`(now `terraced`): Terraced (including end-terrace)

As a first step, we can explore the relationship between the two variables. The code below illustrates the use of the `ggpairs` of the `GGally` library, which provides additional and more complex plots on top of the `ggplot2` framework.

```
# install.packages("GGally")
library(GGally)

leicester_dwellings %>%
  dplyr::select(perc_semidetached, perc_terraced) %>%
  GGally::ggpairs(
    upper = list(continuous = wrap(ggally_cor, method = "kendall")),
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```



The scatterplot above seems to indicate that at least three clusters might exist in the data. One with a lot of semi-detached and few terraced house (top-left corner of the scatterplot); one with a lot of terraced and few semi-detached houses (bottom-right corner of the scatterplot); and one with very few of both

classes (bottom-left corner of the scatterplot).

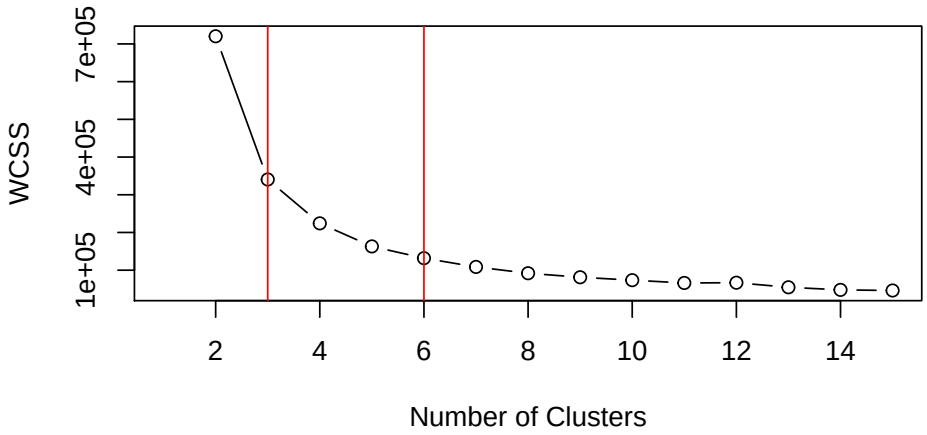
However, there are clearly many OAs which populate the area in between those three groups. So, what is the best approach to cluster those OAs?

The code below illustrates how to create three plots. The first follows the elbow method heuristic. The second and the third take a similar approach but using the silhouette and gap statistic.

```
# Get only the data necessary for testing
data_for_testing <-
  leicester_dwellings %>%
  dplyr::select(perc_semidetached, perc_terraced)

# Calculate WCSS and silhouette
# for k = 2 to 15
# Set up two vectors where to store
# the calculated WCSS and silhouette value
testing_wcss <- rep(NA, 15)
testing_silhouette <- rep(NA, 15)

# for k = 2 to 15
for (testing_k in 2:15){
  # Calculate kmeans
  kmeans_result <-
    stats::kmeans(data_for_testing, centers = testing_k, iter.max = 50)

  # Extract WCSS
  # and save it in the vector
  testing_wcss[testing_k] <- kmeans_result %$% tot.withinss

  # Calculate average silhouette
  # and save it in the vector
  testing_silhouette[testing_k] <-
    kmeans_result %$% cluster %>%
    cluster::silhouette(
      data_for_testing %>% dist()
    ) %>%
    magrittr::extract(, 3) %>% mean()
}

# Calculate the gap statistic using bootstrapping
testing_gap <-
  cluster::clusGap(
    data_for_testing,
    FUN = kmeans,
    K.max = 15, # max number of clusters
```

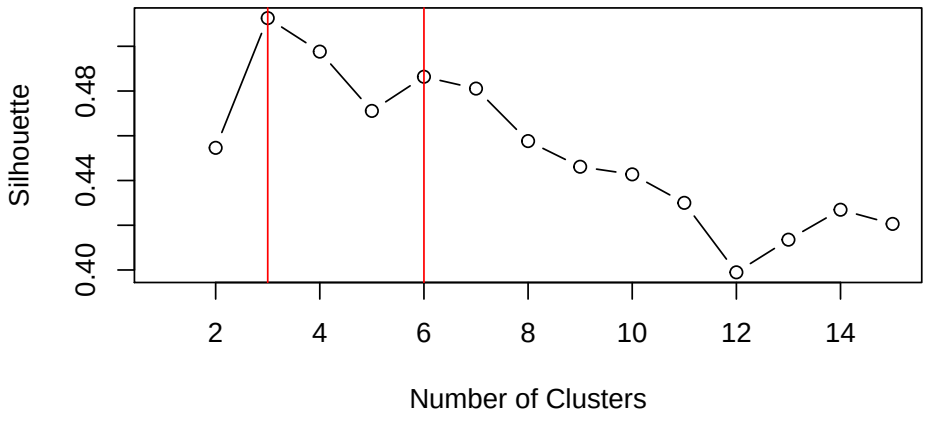
```
    B = 50       # number of samples
  )
```

```
# Plots
plot(2:15, testing_wcss[2:15], type="b", xlab="Number of Clusters",
     ylab="WCSS", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```



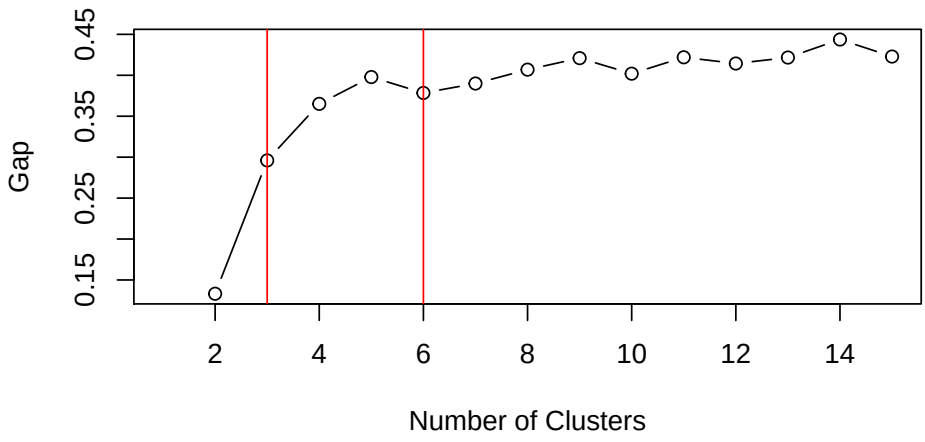

```
## integer(0)
```

```
plot(2:15, testing_silhouette[2:15], type="b", xlab="Number of Clusters",
     ylab="Silhouette", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```




```
## integer(0)
```

```
plot(2:15, testing_gap[["Tab"]][2:15, "gap"], type="b", xlab="Number of Clusters",
     ylab="Gap", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```




```
## integer(0)
```

Based on the WCSS plot, the number of clusters k best fitting the data could range between `k = 3` and `k = 6`, which are all around the inflation point (elbow) of the line. The silhouette plot shows a local maximum at `k = 3` and `k = 6`, which indicates that the observations best fit within their clusters when 3 or 6 clusters are created. The gap statistic steadily increases until it reaches a plateau around `k = 5`. That indicates that clustering improves as we move from 2 to 5 clusters, but the quality doesn't increase as much afterwards. The value for `k = 6` (which is the value suggested by the other two heuristics) is a local minimum, but still the difference with neighbouring values is relatively small.

Overall, the heuristics seem to indicate that `k = 3` could lead to a good clustering result. However, the gap statistic indicates that those three clusters would not be very compact. That is probably due to the fact that the observations at center of the scatterplot seen above are rather uniformally distributed over the space between the three main clusters. As such, chosing `k = 6` clusters might be the best fitting approach in this case. We can then calculate the clusters for `k = 6` as shown below.

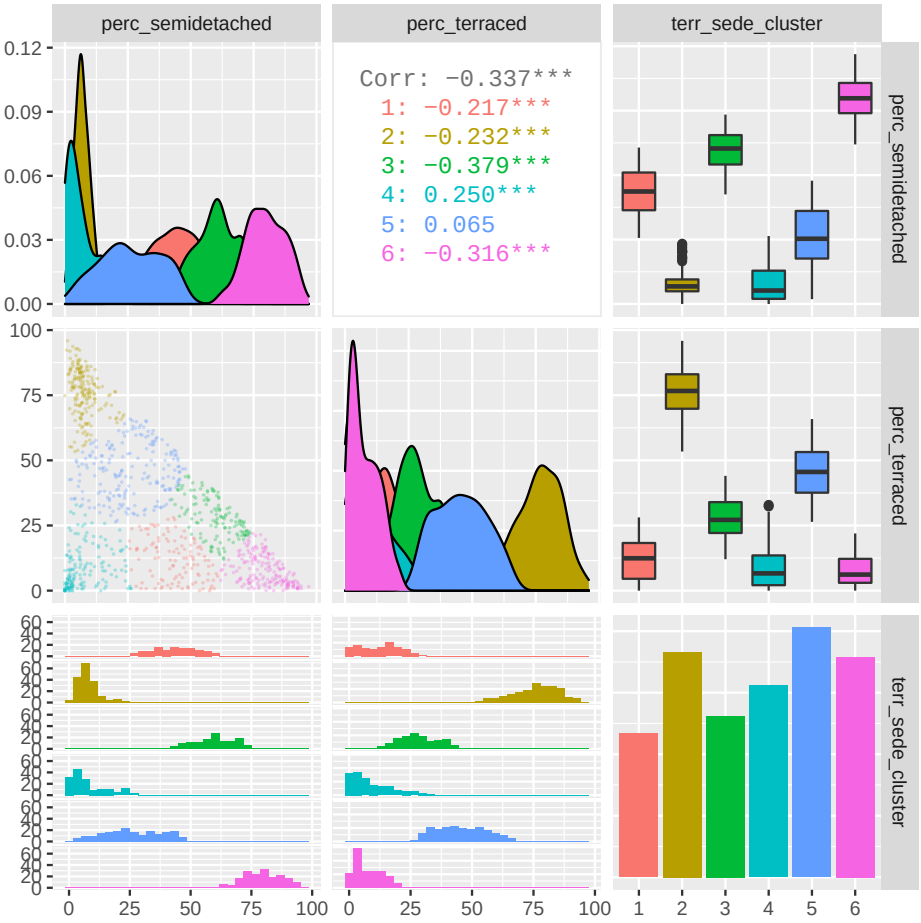
```
terr_sede_kmeans <- leicester_dwellings %>%
  dplyr::select(perc_semidetached, perc_terraced) %>%
  stats::kmeans(centers = 6, iter.max = 50)

leicester_dwellings <-
  leicester_dwellings %>%
  tibble::add_column(
```

```
    terr_sede_cluster = terr_sede_kmeans %$% cluster %>% as.character()
  )
```

10.2.2 Interpreting the clusters

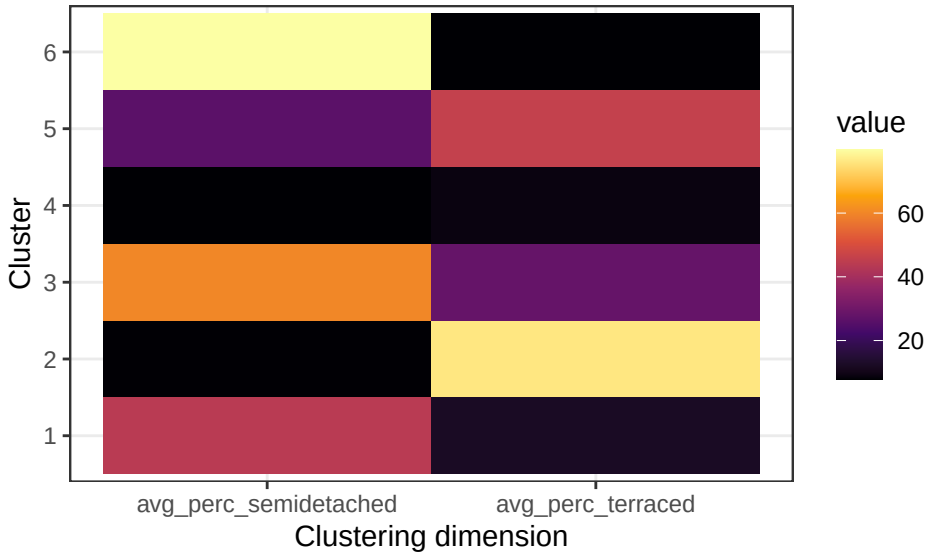
After the clustering has been completed, we can analyse the results through a visual analysis. For instance, we can plot the two original variables, along with the computed clusters as illustrated below.

```
leicester_dwellings %>%
  dplyr::select(perc_semidetached, perc_terraced, terr_sede_cluster) %>%
  GGally::ggpairs(
    mapping = aes(color = terr_sede_cluster),
    upper = list(continuous = wrap(ggally_cor, method = "kendall")),
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```



Another common approach in interpreting the results is to create *heatmaps* for the average values of the variables used in the clustering process for each cluster.

```
leicester_dwellings %>%
  group_by(terr_sede_cluster) %>%
  dplyr::summarise(
    avg_perc_semidetached = mean(perc_semidetached),
    avg_perc_terraced = mean(perc_terraced)
  ) %>%
  dplyr::select(terr_sede_cluster, avg_perc_semidetached, avg_perc_terraced) %>%
  tidyr::pivot_longer(
    cols = -terr_sede_cluster,
    names_to = "clustering_dimension",
    values_to = "value"
  ) %>%
  ggplot2::ggplot(
    aes(
      x = clustering_dimension,
      y = terr_sede_cluster
    )
  ) +
  ggplot2::geom_tile(aes(fill = value)) +
  ggplot2::xlab("Clustering dimension") +
  ggplot2::ylab("Cluster") +
  ggplot2::scale_fill_viridis_c(option = "inferno") +
  ggplot2::theme_bw()
```



The plot above, clearly illustrates how cluster 6 has a high percentage of semi-

detached houses and a low percentages of terraced houses. Cluster 2 has a high percentage of terraced houses and a low percentages of semi-detached houses. Cluster 4 has a low percentage of both semi-detached and terraced houses. Those are the three clusters that we first identified from the first scatterplot above.

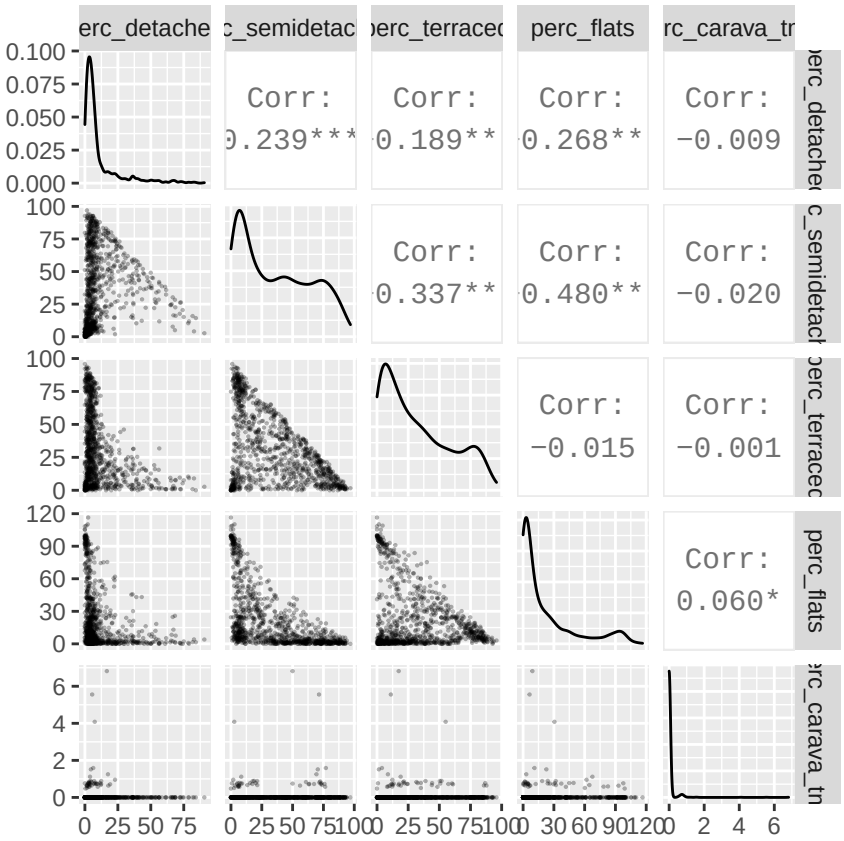
Moreover, the clustering process identifies cluster 1, which includes similar percentages of semi-detached and terraced houses; as well as cluster 5, including mostly semi-detached but also some terraced houses, and cluster 3, including mostly terraced but also some semi-detached houses.

10.2.3 A geodemographic of dwelling types

The case study above is useful as a simple example of how to create a geodemographic classification, but possibly not the most interesting analysis due to the limited number of variables used. In this section, we explore the creation of a geodemographic of dwelling types in the city of Leicester, using all five variables available in the original dataset.

As before we can start from a visual analysis of the data. The relationship between the different variables generally resembles the one seen between semi-detached and terraced houses seen in the example above, except for caravans and mobile or temporary structures, which seem fairly rare in Leicester. No variable seems to be particularly well correlated to any other. Thus all variables should be included in the classification, as they all have the potential to contribute relevant information.

```
leicester_dwellings %>%
  dplyr::select(perc_detached:perc_carava_tmp) %>%
  GGally::ggpairs(
    upper = list(continuous = wrap(ggally_cor, method = "kendall")),
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```

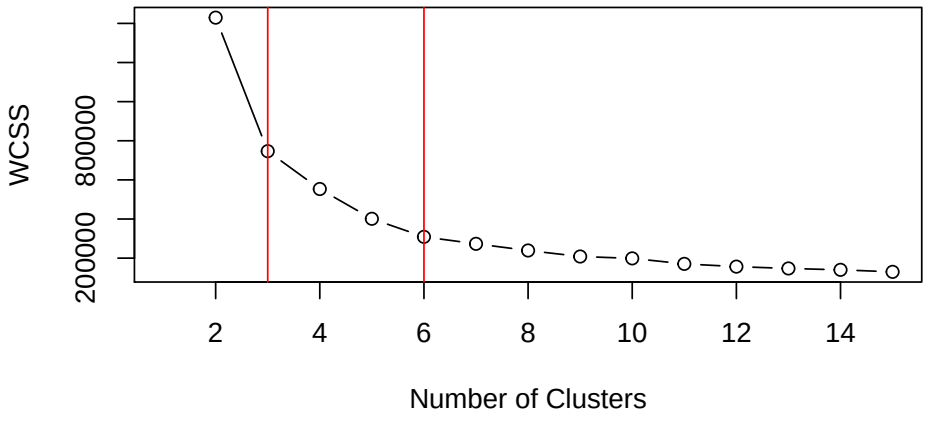


In order to identify the number of clusters `k` which best fits the data, we can use the elbow method, along with the silhouette and gap statistic measures, as seen in the previous example. The only difference is that in this case `data_for_testing` will include all five variables.

```
# Data for elbow method
data_for_testing <-
  leicester_dwellings %>%
  dplyr::select(perc_detached:perc_carava_tmp)

# Calculate WCSS and silhouette
# for k = 2 to 15
# Set up two vectors where to store
# the calculated WCSS and silhouette value
testing_wcss <- rep(NA, 15)
testing_silhouette <- rep(NA, 15)

# for k = 2 to 15
for (testing_k in 2:15){
```

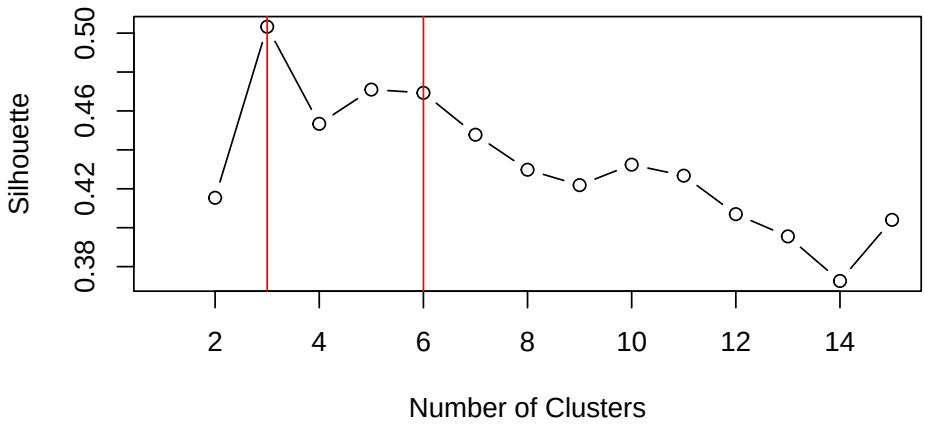
```
  # Calculate kmeans
  kmeans_result <-
    stats::kmeans(data_for_testing, centers = testing_k, iter.max = 50)

  # Extract WCSS
  # and save it in the vector
  testing_wcss[testing_k] <- kmeans_result %$% tot.withinss

  # Calculate average silhouette
  # and save it in the vector
  testing_silhouette[testing_k] <-
    kmeans_result %$% cluster %>%
    cluster::silhouette(
      data_for_testing %>% dist()
    ) %>%
    magrittr::extract(, 3) %>% mean()
}

# Calculate the gap statistic using bootstrapping
testing_gap <-
  cluster::clusGap(data_for_testing, FUN = kmeans,
    K.max = 15, B = 50
  )
```

```
# Plots
plot(2:15, testing_wcss[2:15], type="b", xlab="Number of Clusters",
     ylab="WCSS", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```



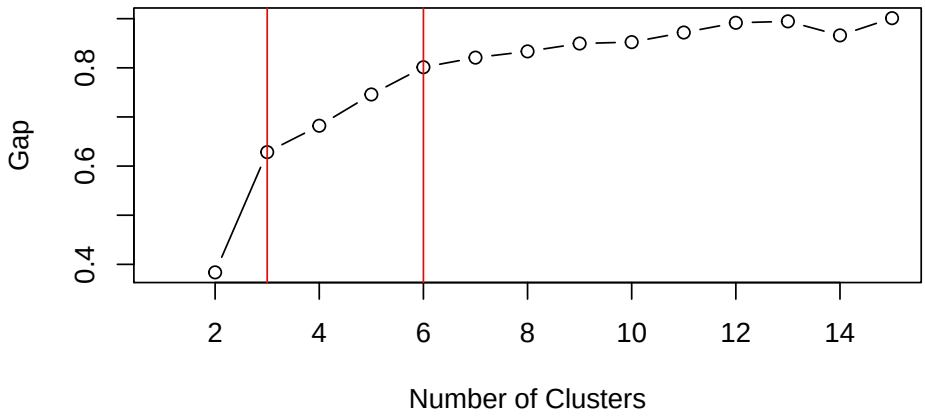

```
## integer(0)
```

```
plot(2:15, testing_silhouette[2:15], type="b", xlab="Number of Clusters",
     ylab="Silhouette", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```




```
## integer(0)
```

```
plot(2:15, testing_gap[["Tab"]][2:15, "gap"], type="b", xlab="Number of Clusters",
     ylab="Gap", xlim=c(1,15)) +
abline(v = 3, col = "red") +
abline(v = 6, col = "red")
```




```
## integer(0)
```

As in the previous example, the elbow method (i.e., WCSS), silhouette and gap statistic seem to indicate that `k = 3` or `k = 6` might be the best choice. Let's see what the result is when choosing `k = 6`.

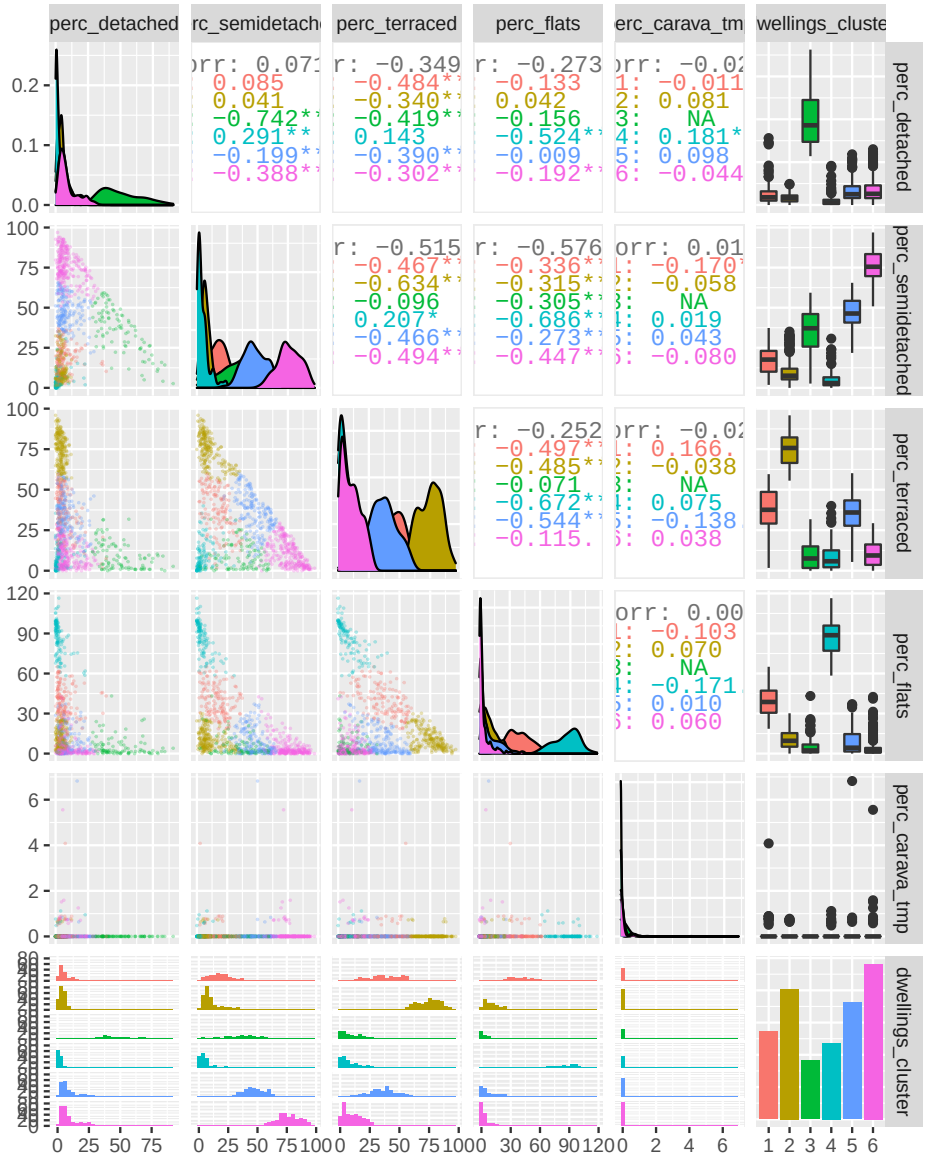
```
dwellings_kmeans <- leicester_dwellings %>%
  dplyr::select(perc_detached:perc_carava_tmp) %>%
```

```
  stats::kmeans(
    centers = 6,
    iter.max = 50
  )

leicester_dwellings <-
  leicester_dwellings %>%
  tibble::add_column(
    dwellings_cluster =
      dwellings_kmeans %$%
        cluster %>%
          as.character()
  )
```

10.2.4 Interpreting the clusters

A first exploratory plot of the clusters seems to reveal clusters that closely resemble those seen in the first example above.

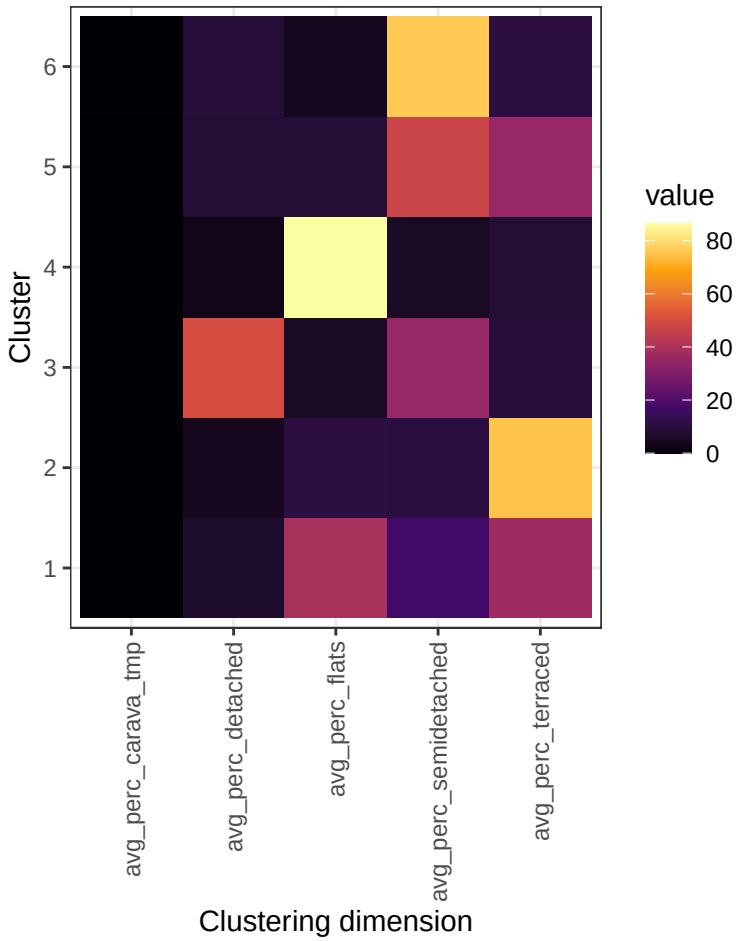
```
leicester_dwellings %>%
  dplyr::select(perc_detached:perc_carava_tmp, dwellings_cluster) %>%
  GGally::ggpairs(
    mapping = aes(color = dwellings_cluster),
    lower = list(continuous = wrap("points", alpha = 0.3, size=0.1))
  )
```


perc_detached
rc_semidetach
perc_terraced
perc_flats
erc_carava_tm
wellings_cluste
perc_detached
perc_semidetached
perc_terraced
perc_flats
perc_carava_tmp
dwellings_cluster

As in the previous example, we can use an *"heatmap"* plot to explore how the clusters are characterised by the variables used in the clustering process (see also Exercise 414.1.1 below).

```
dwellings_cluster_avgs <-
  leicester_dwellings %>%
  group_by(dwellings_cluster) %>%
  dplyr::summarise(
    dplyr::across(
      perc_detached:perc_carava_tmp,
      mean
    )
  ) %>%
  # rename columns
  dplyr::rename_with(
    function(x){ paste0("avg_", x) },
    perc_detached:perc_carava_tmp
  )

dwellings_cluster_avgs %>%
  tidyr::pivot_longer(
    cols = -dwellings_cluster,
    names_to = "clustering_dimension",
    values_to = "value"
  )  %>%
  ggplot2::ggplot(
    aes(
      x = clustering_dimension,
      y = dwellings_cluster
    )
  ) +
  ggplot2::geom_tile(
    aes(
      fill = value
    )
  ) +
  ggplot2::xlab("Clustering dimension") +
  ggplot2::ylab("Cluster") +
  ggplot2::scale_fill_viridis_c(option = "inferno") +
  ggplot2::theme_bw() +
  ggplot2::theme(
    axis.text.x =
      element_text(
        angle = 90,
        vjust = 0.5,
        hjust=1
```

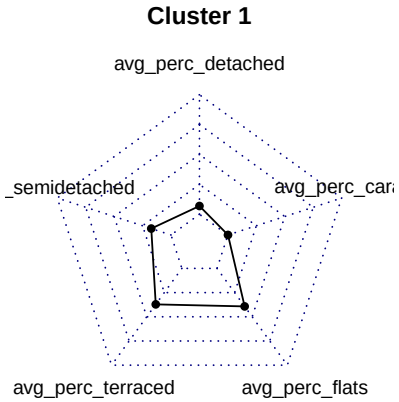
```
    )
  )
```

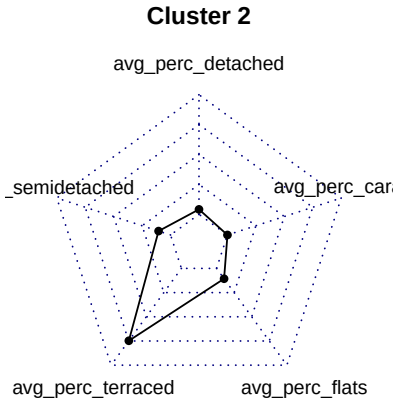


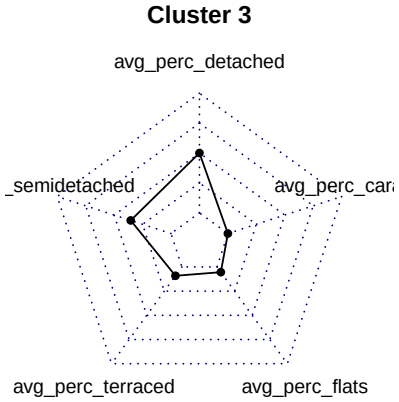
Another very common approach to explore the characteristics of the clusters created through k-means for the geodemongraphic classification is to use radar charts (also known as spider charts, web charts or polar charts), which can be created in R using a number of libraries, including the `radarchart` of the `fmsb` library.

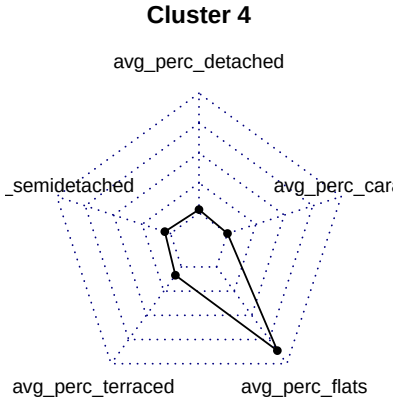
```
# install.packages("fmsb")
library(fmsb)

par(mar=rep(3,4))
par(mfrow=c(3,2))

for(cluster_number in 1:6){
```

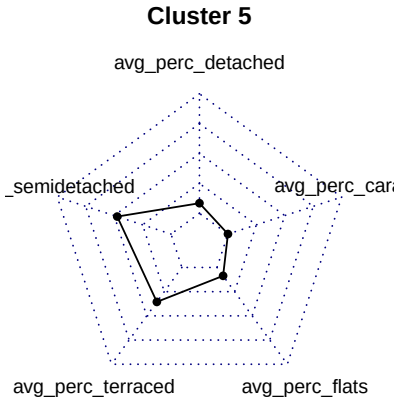
```
  rbind (
    # The radar chart requires a maximum and a minimum row
    # before the actual data
    rep(100, 5), # max 100% for 5 variables
    rep(0, 5),   # min 0% for 5 variables
    dwellings_cluster_avgs %>%
      dplyr::filter(dwellings_cluster == cluster_number) %>%
      dplyr::select(-dwellings_cluster) %>%
      as.data.frame()
    ) %>%
    fmsb::radarchart(title = paste("Cluster", cluster_number))
}
```

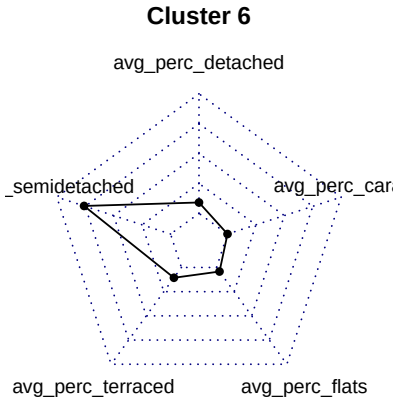

Cluster 1
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats


Cluster 2
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats


Cluster 3
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats


Cluster 4
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats

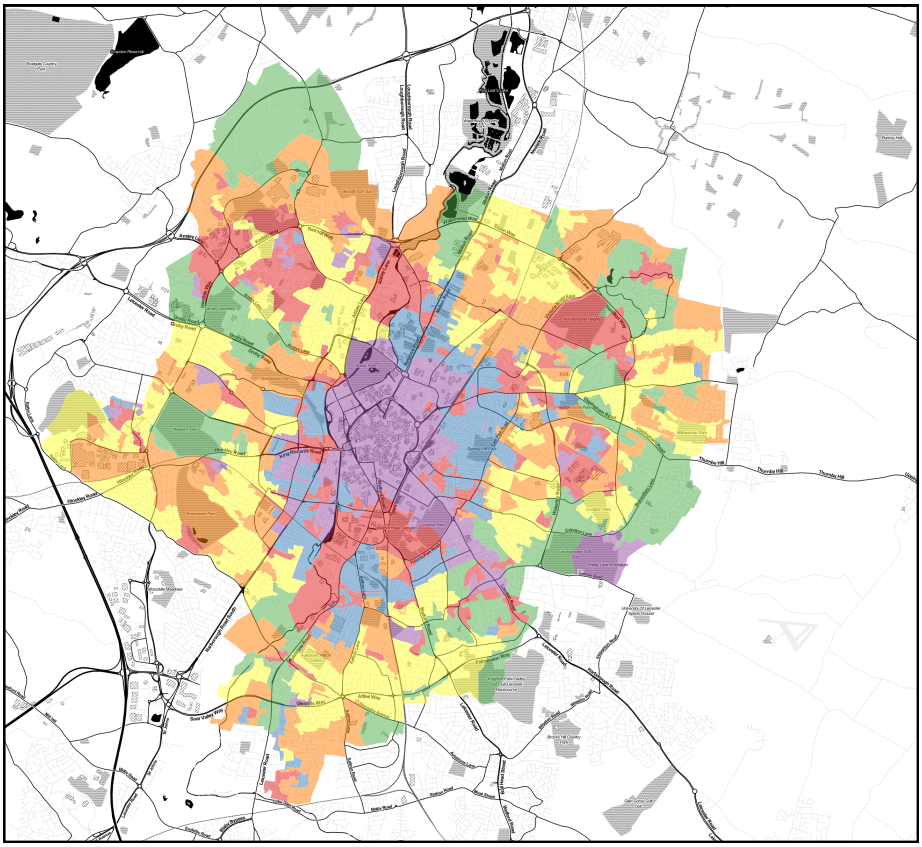

Cluster 5
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats


Cluster 6
avg_perc_detached
_semidetached
avg_perc_terraced
avg_perc_flats

The radar charts are very effective in visualising the values for multiple varaibles, as long as the variables are all of similar type, value and range. In this case, as all values are percentages, radar chart are very effective in illustrating which variables have particularly high averages in each cluster.

Finally, we can map the cluster cartographically to analyse their spatial distribution.

Geodemographic classification of dwellings



Source: CDRC 2011 OAC Geodata Pack by the ESRC Consumer Data Research Centre; Contains National Statistics data Crown copyright and database right 2015; Contains Ordnance Survey data Crown copyright and database right 2015. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

10.3 Exercise 414.1

Question 414.1.1: Based on the *"heatmap"*, radar charts and map created in the example above, how would you characterise the five clusters? How would you name them?.

Question 414.1.2: Create a geodemographic classification using the data seen in the second example above, but creating `k = 9` clusters.

Question 414.1.3: Create a geodemographic classification for the city of Leicester based on the presence of peoples in the different age groups included in the `2011_OAC_Raw_uVariables_Leicester.csv` dataset (`u007` to `u019`).