

AN INTRODUCTION TO R FOR POLICY ANALYSIS

(Week 2)



COURSE SCHEDULE:

Webinar each Monday @ 10am (AEST) ~ 20 July to 17 August

- ~~Week 1 ~ **An introduction to programming**, the R language and Rstudio~~
- Week 2 ~ **Doing stuff with data** - Importing, Exploring and Summarizing Data
- Week 3 ~ **Making pixels pretty** (1) - Data Cleaning, Merging and Basic Visualization
- Week 4 ~ **Making pixels pretty** (2) - Data Cleaning and Visualization
- Week 5 ~ **Bringing together the R programming pipeline**

POLL: COMPLETED SWIRL EXERCISES?

- Building Blocks
- Sequences of Numbers
- Missing Values



INTRODUCTIONS: BARBARA PERRY

- Applied statistician with a Masters in Animal Science and a Masters in Arts (Teaching)
- Experienced laboratory technologist, statistical analysis and teacher
- Most recently working as a statistical analyst for Government in the State of Nebraska where I provide advice to the leadership of, and stakeholders to, the Department of Health and Human Services
- In practice this requires the use of a combination of tools such as Access, Excel and R to translate multiple data sources into actionable insights

WORKSHOP 2 OUTLINE:

- Refresher: A brief reminder of some basics
- The Tidyverse
- dplyr: Useful Data Plier Verbs
- Titanic analysis: using dplyr
- Policy Scenario: Using UK Census data to examine housing ownership

REFRESHER : THE BASICS

THE BASICS:

- **R console** - where R does stuff
- **R studio** – a program to help us use R
- **Scripts** – Text file with our ‘analysis recipe’
- **Packages** – Groups of commands that do stuff
- **Environment** – Where R stores stuff it works with
- **Working directory** – where your project is
- **Base R** – R without loading packages

RSTUDIO HELPS US USE R:

The screenshot shows the RStudio interface with four panels highlighted by text boxes:

- SCRIPTS**
(analysis recipes)
- ENVIRONMENT**
(where data and stuff is stored)
- R CONSOLE**
(where we tell R what to do)
- EXPLORER**
(viewer for plots/help/files)

The R Script editor shows the following code:

```
1 normaldata<- rnorm(1000)
2 hist(normaldata, main="Example Histogram")
3
```

The Environment pane shows the following values:

Values	
normaldata	num [1:1000] -0.2111 -1.1145 0.0408 1.1039 ...
obj	List of 1

The Console pane shows the following output:

```
uting
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative effort of many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R in publications.

Type 'demo()' for some demos, 'help()' for on-line help,
or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]
```

The Explorer pane shows a histogram titled "Example Histogram" with the x-axis labeled "normaldata" and the y-axis labeled "Frequency". The histogram shows a distribution of values ranging from approximately -2 to 4.

TYPICAL STEPS IN R ANALYSIS:

- **#Set-up** - the working directory ----
- **#Load** the necessary packages ----
- **#Import**, Explore + Clean the data ----
- **#Explain** - analyse, understand and visualize ----
- **#Save** - the results ----

(there is a template in the dropbox which follows this structure)

ELEMENTS (WE CARE ABOUT)

Element types:

- **Numeric** - 1.2, 45.0, 21.6
- **Character** - "Ben" , "Sheryl", "Bazza"
- **Dates** – '23/12/85'
- **Logical** - TRUE, FALSE
- **Factors** – numbers and text (used for labelling)

OBJECT TYPES

- **Vectors** ~ one column one type of element
- **Matrices** ~ multiple columns of one element type
- **Data Frames (or Tibbles)** ~ multiples columns of any element type
- **Lists** ~ Collection of objects (eg a bunch of data frames)

(Objects are collections of elements)

eg: `stuff <- c(31,42,13)`

SUBSETTING DATA

Vectors (use `x[row]`)

- `vector_object<-(1:100)`
- 33rd item: `vector_object[33]`

Data frames `x[row,columns]`

- `data_object<-data.frame(1:100, 100:1,1:100*5)`
- 33rd item, in the 2nd column: `Data_object[33,2]`
- A range of rows and exclude a column:
`data_object[33:34, -3]`

OR Select columns directly with `$` via data frames

Eg `data_object$column_name` and `data_object$column_name[33]`

TESTING CONDITIONS WITH LOGIC

- Less than: <
- Greater than: >
- Equal to: == or Not equal to !=
- And: &
- Or: |
- Not: ! (eg, not equal !=, not greater than !> and not less than !<)

These are useful for extracting and summarizing data based on logical conditions. The result will be a logical vector.

*For instance: 1:5>2 results in:
FALSE FALSE TRUE TRUE TRUE*

BASE R AND THE TIDYVERSE

COMMANDS | FUNCTIONS | PACKAGES

R functions/commands: actions we tell the computer to do

- `print("Get off my lawn!")`
- `Average(1:1000)`

Using them requires that we feed the right ingredients to work:

- `write.csv(object_name, "filename_to_write.csv")`

R Packages = a bunch of functions stuck together

- **Base R** = packages that come with R by default
- **Tidyverse** = collection of packages that work together to help do most common data-related tasks

(see: tidyverse.org)

PACKAGES

Packages can be installed by typing:

```
install.packages('package_name')
```

To use a package we then have to load it by typing:

```
library(package_name)
```


THE TIDYVERSE

Tidverse = packages you're likely to use in
everyday data analysis.

Intuitive grammar and logic make R
simpler

Swirl exercises focus on Base R

CORE TIDYVERSE:

Get/Save
Data

readr

readxl

haven

Data
Cleaning

lubridate

stringr

tidyr

Analysis and
Data viz

dplyr

ggplot2

TIDYVERSE DATA FRAMES ~ TIBBLES

The tidyverse is a 'modern' version of the data.frame designed to do what is useful and drop what such as by:

- Never automatically changing the element type Not modifying column names
- Displaying more useful information about the data
- Not 'guessing' which column you *might* want when using \$

Generally, tidyverse commands that replicate base R commands use '_' instead of '.'

`read_csv()` **instead of** `read.csv()`

Note: Some commands won't work with tibbles, but you can convert them to data frames using `as_data_frame()`

TIDY DATA

TIDY DATA:

Data being 'Tidy' is a common theme in R programming, requiring that:

- each variable is a column
- each observation a row; and
- each type of observational unit a separate table

(the Tidyverse packages also tend to work better with 'tidy' data)

TIDY DATA:

Tidy data looks like this:

Vehicle	Efficiency	Power	Year
Mazda RX4	21	110	2001
Mazda RX4 Wag	21	110	2015
Merc 230	22.8	95	2015
Merc 280	19.2	123	1983
Datsun 710	22.8	93	2017

- ✓ Each variable is a column.
- ✓ Each observation a row.
- ✓ Each experiments/survey a separate table

SUMMARIZING DATA WITH DPLYR

DPLYR VERBS

dplyr = 'Data Plier'. Provides a grammar for basic data manipulation:

- **select** – pick the columns you want
- **filter** – choose the relevant observations
- **mutate** – create new variables
- **group_by** – define groups you're interested in for summarise
- **summarise** – create summary statistics
- **arrange** – sort your data

DPLYR: USEFUL SUMMARY FUNCTIONS

- **Center:** `mean()`, `median()`
- **Spread:** `sd()`, `IQR()`, `mad()`
- **Range:** `min()`, `max()`, `quantile()`
- **Position:** `first()`, `last()`, `nth()`,
- **Count:** `n()`, `n_distinct()`
- **Logical:** `any()`, `all()`

*you can see this by seeing the help for summarise using
?`summarise`*

TITANIC ANALYSIS: THE TIDYVERSE WAY

TITANIC SCENARIO

Having reflected on the realism of the 'Titanic' movie, the minister has come back to you with some additional questions about your earlier analysis, asking:

- How much do we *really* know about what happened that night? For instance, does the number of missing values for particular variables suggest we should question the data's accuracy?
- Dropping the 'body' variable and removing all records where we don't know their age, how many people survived by class and sex?
- What was their average age and fare (converted to USD)?
 - (assuming 1 GBP = 1.3 USD)

TITANIC SCENARIO

How much do we *really* know about what happened that night? For instance, does the number of missing values for variables suggest we should question the data's accuracy?

- Set the workspace ~ `setwd()`
- Load the data ~ `read_csv()`
- Examine the number of missing values:
 - **`summary(titanic_data)`** command
 - body and age have a lot of missing values. Is this a feature or bug?
- Examine the missing values by category to figure this out:
 - `table(is.na(titanic_data$body), titanic_data$pclass)`
 - `table(is.na(titanic_data$body), titanic_data$survived)`
 - `table(is.na(titanic_data$age), titanic_data$pclass)`

TITANIC SCENARIO

Dropping the 'body' variable and removing all records where we don't know their age, how many people survived by class and sex?

- Drop the 'body' variable
 - `titanic_data_accurate<- select(titanic_data, -body)`
- Remove observations where the age is missing
 - `titanic_data_accurate <- filter(titanic_data_accurate, is.na(age))`
- Count the number of passengers that survived by class and sex
 - `titanic_data_accurate <- group_by(titanic_data_accurate, pclass, sex, survived)`
 - `titanic_data_accurate <- summarise(titanic_data_accurate, freq=n())`

Note: *we need to assign the results of each line to an object and feed that object to the next line of code for this to work. Eg we need to use the filtered data when assigning groups and creating the summary*

TITANIC SCENARIO

What was their average age and fare (converted to USD)?

(assuming 1 GBP = 1.3 USD)

- **Create a new variable 'fare_usd'**
 - `titanic_data_accurate <- mutate(titanic_data_accurate, fare_usd = fare * 1.3)`
- **Check groups are what we need:**
 - `group(titanic_data_accurate)`
- **Define summaries to calculate:**
 - `summarise(titanic_data_accurate, freq=n(), avg_fare_USD= mean(fare_USD), avg_age= mean(age))`

DPLYR VERBS

dplyr = 'Data Plier'. Provides a grammar for basic data manipulation:

- **select** – pick the columns you want
- **filter** – choose the relevant observations
- **mutate** – create new variables
- **group_by** – define groups you're interested in for summarise
- **summarise** – create summary statistics

THE 'PIPE' OPERATOR %>%

'%>%' or 'pipes' allow results to be sequentially passed through functions:

- **Windows:** CTRL + SHIFT + M
- **Mac:** CMD + Shift + M

With pipes:

```
ingridients %>%  
  mix() %>%  
  pour(into=baking_form) %>%  
  put(into=oven) %>%  
  bake(time=30) %>%  
  slice(pieces=6) %>%  
  eat(1)
```

Without pipes:

```
eat(  
  slice(  
    bake(  
      put(  
        pour(  
          mix(ingridients),  
          into=baking_form),  
          into=oven),  
        time=30),  
        pieces=6),  
    1)
```

Source: <https://twitter.com/dmi3k/status/1191824875842879489>

THE 'PIPE' OPERATOR %>%

In the titanic example:

```
survival_class_sex_and_fare_piped<- titanic_data %>%  
  filter(age != is.na(age)) %>%  
  select(-body) %>%  
  mutate( fare_USD = fare*1.3) %>%  
  group_by(survived_logical, pclass, sex) %>%  
  summarise(freq=n(),  
            avg_fare_USD= mean(fare_USD),  
            avg_age= mean(age))
```

A good way to think about '%>% ' is that it feeds something into what it points to (such as a data, results etc)

POLICY SCENARIO

UK Census Microdata

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 data, ignoring age)
- How does this change if we only focus on Scotland?

POLICY SCENARIO: UK CENSUS DATA

Which of these do we need for this?

- **setwd()** – tell R where the data is
- **library()** – load packages we need
- **read_spss()** _ for importing spss data
- **select()** – pick the columns you want
- **filter()** – choose the relevant observations
- **mutate()** – create new variables
- **group_by()** – define groups you're interested in for summarise
- **summarise()** – create summary statistics
- **arrange()** – sort your data

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 data, ignoring age)
- How does this change if we only focus on Scotland?

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The **% of the population that reside in an 'owner occupied' house in 1981** (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 data, ignoring age)
- How does this change if we only focus on Scotland?

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- **The % of the population that reside in an 'owner occupied' house in 1981 according to age** (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 data, ignoring age)
- How does this change if we only focus on Scotland?

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (**comparing 1971 and 1981 results, ignoring age**)
- How does this change if we only focus on Scotland?

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 results, ignoring age)
- How does this change if we **only focus on Scotland?**

POLICY SCENARIO: UK CENSUS DATA

Perceived decreases in housing affordability have placed the Minister of housing under an intense level of pressure from the public.

Excluding those that are not present residents (popbase), the minister has asked you to undertake some quick analysis on census microdata to determine:

- The % of the population that reside in an 'owner occupied' house in 1981 (dvtenureo)
- The % of the population that reside in an 'owner occupied' house in 1981 according to age (dvtenureo x ageo)
- Whether the recent declines in the % of owner occupied housing just reflects a long-term trend (comparing 1971 and 1981 data, ignoring age)
- How does this change if we only focus on Scotland?

SUGGESTED SWIRL EXERCISES

Before next week complete

- ~~Building Blocks~~
- ~~Sequences of Numbers~~
- ~~Missing Values~~
- **Subsetting Vectors**
- **Matrices and Data Frames**
- **Looking at Data**
- **Dates and Times**



END OF WORKSHOP 2

PLEASE FILL IN THE ONLINE EVALUATION

(link will be shared in the chat)

AN INTRODUCTION TO R FOR POLICY ANALYSIS

