R for Excel users workshop

Sean Manzi[1]

[1]NIHR South West Peninsula Applied Research Collaboration (PenARC)
University of Exeter

NHS-R Conference, 2019

Aims of the workshop

This workshop will help you to understand:

- How to develop generic R scripts to replace manual Excel based analyses
- Which types of tasks are most appropriate for automation

Workshop outline

- Identifying Excel tasks for translation into R
- Breaking down the components of an Excel analysis
- Building a first-cut non-generic R script
- Making your R script generic
- Conclusions, hints and tips

NIHR | Applied Research Collaboration
South West Peninsula

Identifying Excel tasks for translation into R

The regularity check:

- How regularly do you carry out the task?
- Is the same analysis always carried out?
- Is the shape of the data always the same?

NIHR | Applied Research Collaboration
South West Peninsula

Identifying Excel tasks for translation into R

- What routine tasks do you undertake?

Identifying Excel tasks for translation into R

- What routine tasks do you undertake?
- Which of these tasks do you think could most readily be automated?

NIHR | Applied Research Collaboration
South West Peninsula

Identifying Excel tasks for translation into R

- What routine tasks do you undertake?
- Which of these tasks do you think could most readily be automated?
- Do you automate any of your analyses already and if so how?

NIHR | Applied Research Collaboration South West Peninsula

Identifying Excel tasks for translation into R

What are we looking for?

- regularly conducted tasks
- consistantly shaped data extracts
- consistant analyses

NIHR | Applied Research Collaboration
South West Peninsula

Identifying Excel tasks for translation into R

What are we looking for?

- regularly conducted tasks
- consistantly shaped data extracts
- consistant analyses

The likely suspects

- Performance reporting
- Operational reporting
- Board reports

NIHR | Applied Research Collaboration
South West Peninsula

Identifying Excel tasks for translation into R

Where do I start?

- Choose something simple
- Choose something well defined
- Choose something useful

Breaking down an Excel analysis

Todays scenario

You are an analyst in a healthcare organisation. Every month you have to produce a report on the number of patients on the waiting list for service X and the patient waiting times for the two teams that provide service X. The data extract that you use contains the; ID, age, referral date, treatment date and the name of the team to which the patient has been referred. You know there are always data quality issues so you have to check for duplications, missing data and incorrectly entered data.
The report you have to build requires you to produce descriptive statistics tables and a series of graphs at different levels of detail for the last two years. Overall patient numbers, overall wait times, frequency of wait times, comparison of overall wait times by team, monthly comparison of wait times by team, monthly comparison of patient numbers by team.

NIHR | Applied Research Collaboration South West Peninsula

Breaking down an Excel analysis

What are the different stages of the analysis process that you think you will need to go through?

NIHR | Applied Research Collaboration
South West Peninsula

Breaking down an Excel analysis

The analysis process

- Identify the subject and aim of the analysis
- Raw data input
- Check and clean the data
- Transform the data
- Analyse the data
- Visualise the analysis
- Interpret the results

Sense check (verify) everything throughout the analysis process

NIHR | Applied Research Collaboration
South West Peninsula

Breaking down an Excel analysis

Todays scenario

You are an analyst in a healthcare organisation. Every month you have to produce a report on the **number of patients** on the waiting list for service X and the patient **waiting times** for the **two teams** that provide service X. The data extract that you use contains the; **ID, age, referral date, treatment date and the name of the team to which the patient has been referred**. You know there are always data quality issues so you have to check for **duplications, missing data and incorrectly entered data**.

The report you have to build requires you to produce **descriptive statistics tables** and a **series of graphs** at different levels of detail for the last two years. **Overall patient numbers, overall wait times, frequency of wait times, comparison of overall wait times by team, monthly comparison of wait times by team, monthly comparison of patient numbers by team**.

NIHR | Applied Research Collaboration
South West Peninsula

Building a first-cut non-generic R script

Your task: In front of you is an envelope with code excerpts and descriptions of what each code excerpt does. Order the code excerpts and descriptions to create a working script. The R file 'analysisScriptV1_Muddled.R' contains these same code excerpts and descriptions. Reorder the code excerpts in the R file to test if you have re-constructed the code in the correct order.

Think about the analysis process that we have just described and use this to help you re-construct the code.

Each code excerpt has a corresponding description. The code excerpts are labelled with numbers and the descriptions are labelled with letters.

Task 1: The answers

- F − 4
- E − 11
- L − 6
- A − 16
- C − 9
- M − 2
- J − 13
- O − 8

- H − 7
- N − 5
- K − 14
- G − 10
- I − 1
- B − 3
- D − 15
- P − 12

NIHR | Applied Research Collaboration
South West Peninsula

Making your R script generic

Your task: You have another envelope of code excerpts in front of
you. This time they are a generic version of the first script.
Re-construct the code side-by-side with the first script matching
the non-generic and generic aspects. Then re-construct the generic
script found in the file 'analysisScriptV2_Muddled.R' to see if you
have re-constructed the script correctly.
Look at the different construction of the non-generic and the
generic scripts. What do you think are the important differences?

NIHR | Applied Research Collaboration
South West Peninsula

Task 2: The answers

- B – 7
- C – 13
- G – 2
- I – 15
- F – 3
- K – 9
- M – 11
- L – 14
- H – 1

- E – 18
- Q – 12
- A – 17
- D – 5
- P – 6
- R – 16
- J – 8
- N – 10
- O – 4

NIHR | Applied Research Collaboration
South West Peninsula

Making your script generic

What were the main differences between the code in task 1 and the code in task 2?

NIHR | Applied Research Collaboration
South West Peninsula

Making your script generic

- Make use of user-defined functions
- Remove direct calls to specific columns or rows
- Remove any specific value entries where possible, create the code to automate the identification of these values
- Standardise the shape of the input data or make the script flexible
- Add feedback, checks and warnings to aid the user in seeing what is happening or has gone wrong

NIHR | Applied Research Collaboration South West Peninsula

Conclusions, hints and tips

Pick tasks that are suitable for automation

- regularly conducted tasks
- consistantly shaped data extracts
- consistant analyses
- Choose something simple
- Choose something well defined
- Choose something useful

NIHR | Applied Research Collaboration
South West Peninsula

Conclusions, hints and tips

Follow the analysis process

- Identify the subject and aim of the analysis
- Raw data input
- Check and clean the data
- Transform the data
- Analyse the data
- Visualise the analysis
- Interpret the results

Sense check (verify) everything throughout the analysis process

NIHR | Applied Research Collaboration
South West Peninsula

Conclusions, hints and tips

When writing the program

- Start with a non-generic script
- Make the non-generic script generic
- The program will evolve over time
- Comment your script well
- Share and re-use parts of your script
- Keep on learning

NIHR | Applied Research Collaboration
South West Peninsula

Thank you and I hope you had fun while learning

Check out www.rforhealthcare.org for all of the workshop
materials and a variety of materials to help you to get started
with R