Reproducible Analytical Pipelines & their value in fundamental Data Science

Jeroen Minderman

Senior Data Scientist

23 January 2024

Data Science Campus

Fundamental needs for Data Science in NSOs

1. **Skills**
   - Data literacy
   - Programming literacy
   - Following/building Good Practice

2. **Buy-in**
3. **Resource** (~ buy-in!)

RAP in Data Science: efficiencies

- Reproducible Analytical Pipelines (RAPs)
  - Formalised process for automation of analyses
  - Minimise manual steps, maximise transparency & reproducibility

- Improving *quality*, *trust*, *business continuity*
- Create ***efficiency***: saving resource

https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/

RAP in Data Science: a blueprint for skills

- Data Science ≠ RAP; RAP alone ≠ Data Science
- BUT; for RAP we need, e.g:
  - Understanding process/scope
  - Programming skills; R/Python
  - Focus on application & impact

RAP in Data Science: a blueprint for skills

- Data Science ≠ RAP; RAP alone ≠ Data Science
- BUT; for RAP we need, e.g:
  - Understanding process/scope
  - Programming skills; R/Python
  - Focus on application & impact

== Fundamentals of data science

+

efficiency gains

=

win-win

RAP mentoring approach

- Data scientist(s) mentoring small groups in partner NSO’s, e.g.
  - Scoping suitable work
  - Flexible & scalable training
  - Support pipeline development

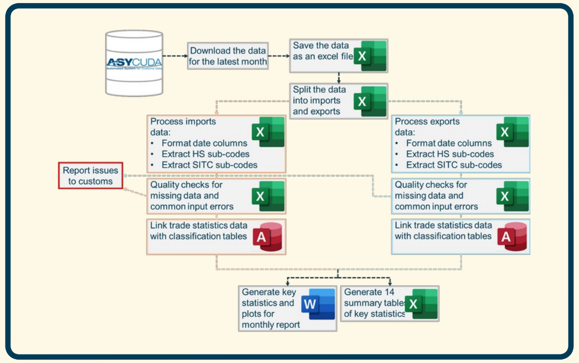
- Longer period with regular check-ins

➔ Focus on ***application*** and ***impact***

Palestine Central Bureau of Statistics (PCBS)

‼ Manual & labour intensive process for trade statistics data

‼ Some R training but not applied



Download the data for the latest month

Save the data as an excel file

Split the data into imports and exports

Process imports data:
- Format date columns
- Extract HS sub-codes
- Extract SITC sub-codes

Process exports data:
- Format date columns
- Extract HS sub-codes
- Extract SITC sub-codes

Report issues to customs

Quality checks for missing data and common input errors

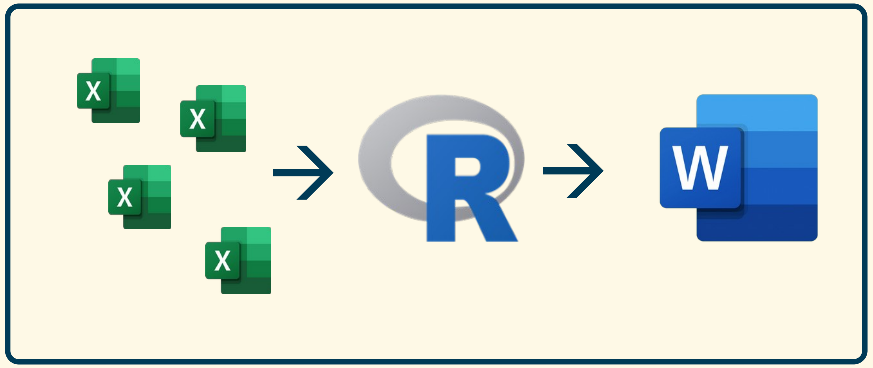
Quality checks for missing data and common input errors

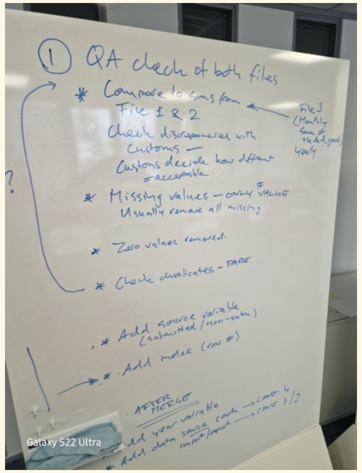
Link trade statistics data with classification tables

Link trade statistics data with classification tables

Generate key statistics and plots for monthly report

Generate 14 summary tables of key statistics




① QA check of both files

* Compare to sums from File 1 & 2 ← File 3 (Monthly sum of traded goods) Yearly

Check discrepancies with Customs — Customs decide how different = acceptable

* Missing values — ONLY IF VALUE
Usually remove all missing

* Zero values removed.

* Check duplicates - RARE

* Add source variable (submitted / non-subm)

* Add index (row #)

AFTER MERGE

Add year variable

* Add data source code → CONST. 4
import/export → CONST. 1/2

✓ Confidence in R, version control → scope for wider DS work

✓ Start on RAP → increased quality & transparency

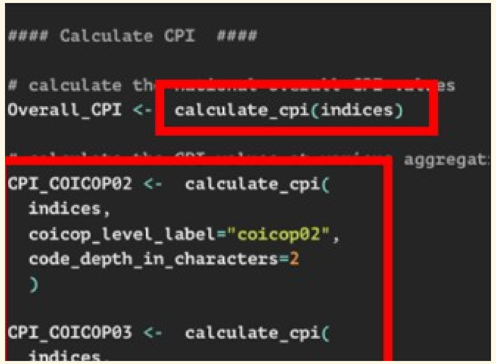
✓ Wider benefits: support overall process improvements

Ghana Statistics Service (GSS)

[blog link]

‼ Skilled in R, part automated process for CPI production

‼ Monthly reporting was done manually

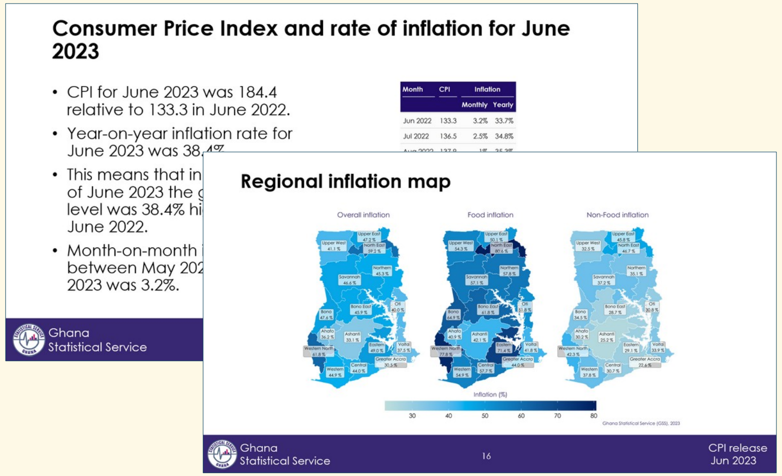


```
#### Calculate CPI  ####

# calculate th                                  es
Overall_CPI <-  calculate_cpi(indices)

                                              aggregat
CPI_COICOP02 <-  calculate_cpi(
  indices,
  coicop_level_label="coicop02",
  code_depth_in_characters=2
  )

CPI_COICOP03 <-  calculate_cpi(
  indices,
```



Consumer Price Index and rate of inflation for June 2023

- CPI for June 2023 was 184.4 relative to 133.3 in June 2022.
- Year-on-year inflation rate for June 2023 was 38.4%
- This means that in of June 2023 the g level was 38.4% hi June 2022.
- Month-on-month between May 202 2023 was 3.2%.

Month	CPI	Inflation	
		Monthly	Yearly
Jun 2022	133.3	3.2%	33.7%
Jul 2022	136.5	2.5%	34.8%
Aug 2022	137.9	1%	35.3%

Ghana Statistical Service

Regional inflation map

Ghana Statistical Service (GSS), 2023

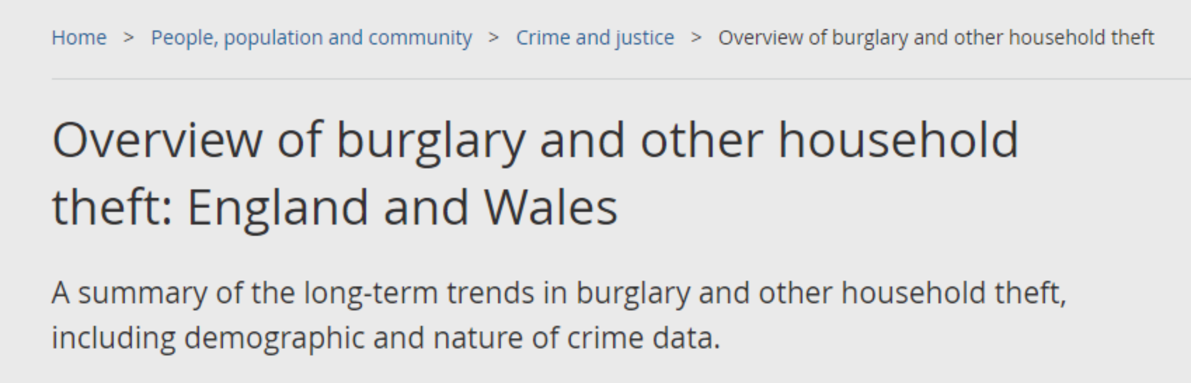
Ghana Statistical Service 16 CPI release Jun 2023

✓ Reproducible report generation → transparency and reliability

✓ Built further confidence in R → mentees now supporting others

RDSA ONS Nature of Crime Automation project

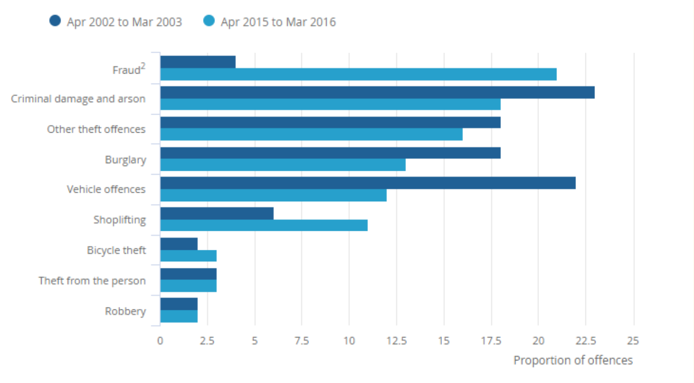
‼ Complex and time-consuming/error prone existing process



Home > People, population and community > Crime and justice > Overview of burglary and other household theft

Overview of burglary and other household theft: England and Wales

A summary of the long-term trends in burglary and other household theft, including demographic and nature of crime data.

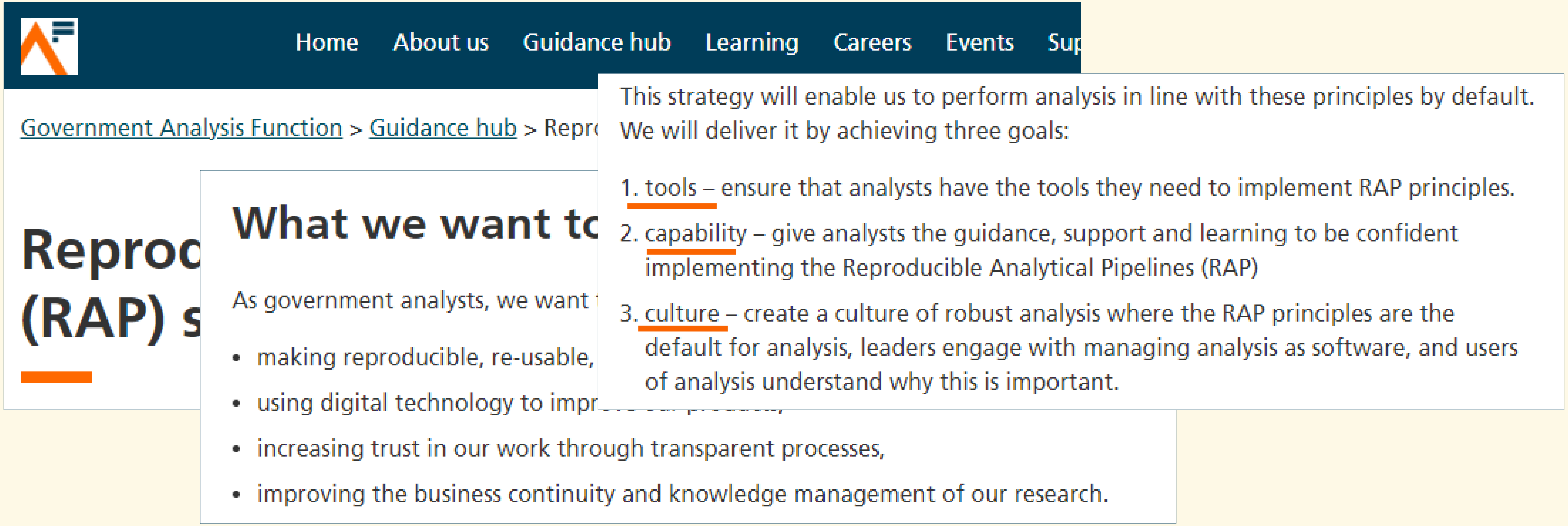


✓ Massive efficiency gains & increased quality / reliability

✓ Working within organisation → able to advise on standardisation of process & reporting

RAP strategy

Standards provide framework for capability & Good Practice



Government Analysis Function > Guidance hub > Repr

Reproc (RAP) s

What we want tc

As government analysts, we want t

- making reproducible, re-usable,
- using digital technology to impr
- increasing trust in our work through transparent processes,
- improving the business continuity and knowledge management of our research.

This strategy will enable us to perform analysis in line with these principles by default. We will deliver it by achieving three goals:

1. tools – ensure that analysts have the tools they need to implement RAP principles.
2. capability – give analysts the guidance, support and learning to be confident implementing the Reproducible Analytical Pipelines (RAP)
3. culture – create a culture of robust analysis where the RAP principles are the default for analysis, leaders engage with managing analysis as software, and users of analysis understand why this is important.

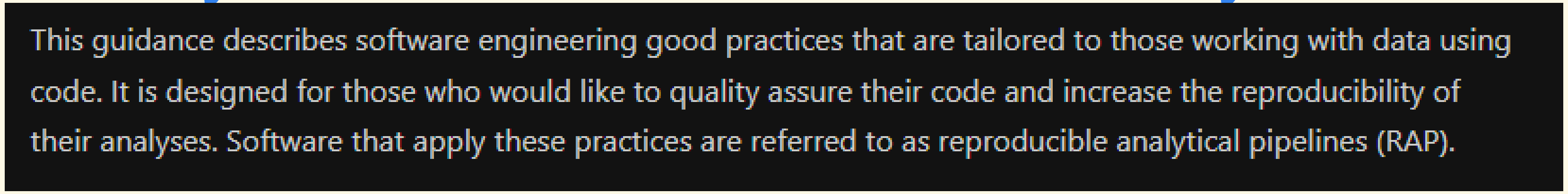
Summary, suggestions & discussion points

- RAP = efficiencies… but also blueprint for DS skills
  - Mentoring is efficient & scalable means to build both
- "Stepping stone" to Pillars 2 and 3

- Notes –
  - Some *initial* skills beneficial; e.g. precede with training courses?
  - Mentor & mentee ***availability is crucial*** (***buy-in***)
    - E.g. ring-fence part of staff time, but plan continued development
  - Take the long view: initial resource cost →→ efficiencies

Resources / links

- ONS Data Science Campus
- UK Analysis Function RAP
- RAP Strategy (UK Analysis Function / ONS)
- RAP case studies
- Using RAP to improve statistics
- Quality Assurance of code for analysis and research



This guidance describes software engineering good practices that are tailored to those working with data using code. It is designed for those who would like to quality assure their code and increase the reproducibility of their analyses. Software that apply these practices are referred to as reproducible analytical pipelines (RAP).