Sharing researcher-generated code and value-added documentation in a Trusted Research Environment

August 2022

Louise Corti, Head Analytical Insights and Impact, ONS

louise.corti@ons.gov.uk

Office for National Statistics

11th Biennial IFC Conference

Overview

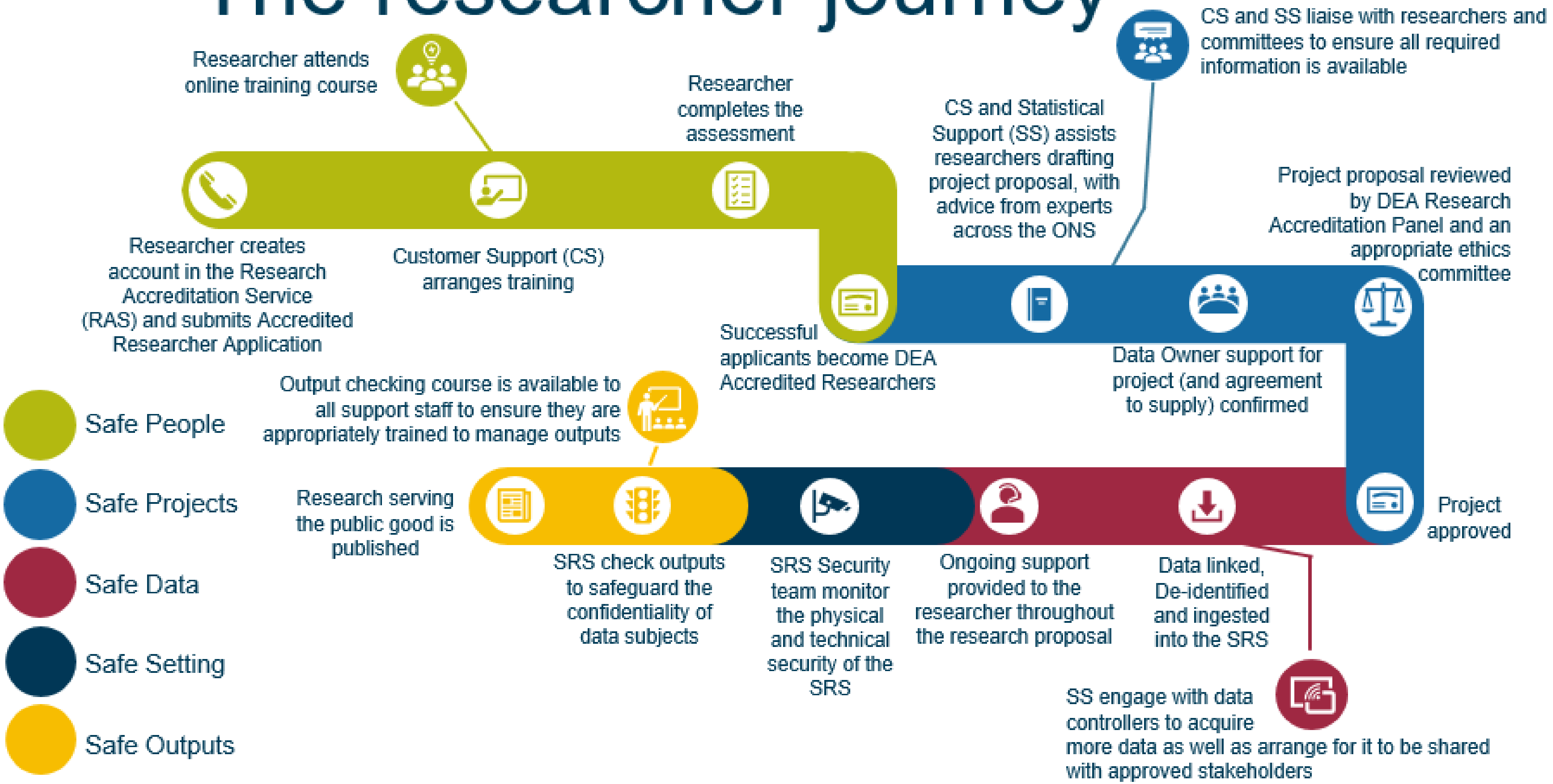
- Introduce researcher code sharing in the UK Secure Research Service (SRS)
- Highlight proposed processes, policies, documentation and support & training activities
- Highlight use cases

Who has shared code with analysts other than with close colleagues?

Why share code?

- ✓Viewed as best practice for demonstrating transparency and accountability in empirical scientific research
- ✓Enables building upon existing code to support the derivation of new variables, avoid recreating complex recoding routines
- ✓Exposes code for peer review /validation/ promotion
- ✓Some journals require underlying code submission
- ❖*Users can feel anxious about exposing their code; consider they don’t have skills or time to write ‘good code’*
- ❖*Users ask if data owners can supply code for derived variables*

The researcher journey



Researcher creates account in the Research Accreditation Service (RAS) and submits Accredited Researcher Application

Customer Support (CS) arranges training

Researcher attends online training course

Researcher completes the assessment

Successful applicants become DEA Accredited Researchers

CS and Statistical Support (SS) assists researchers drafting project proposal, with advice from experts across the ONS

CS and SS liaise with researchers and committees to ensure all required information is available

Data Owner support for project (and agreement to supply) confirmed

Project proposal reviewed by DEA Research Accreditation Panel and an appropriate ethics committee

Project approved

Data linked, De-identified and ingested into the SRS

SS engage with data controllers to acquire more data as well as arrange for it to be shared with approved stakeholders

Ongoing support provided to the researcher throughout the research proposal

SRS Security team monitor the physical and technical security of the SRS

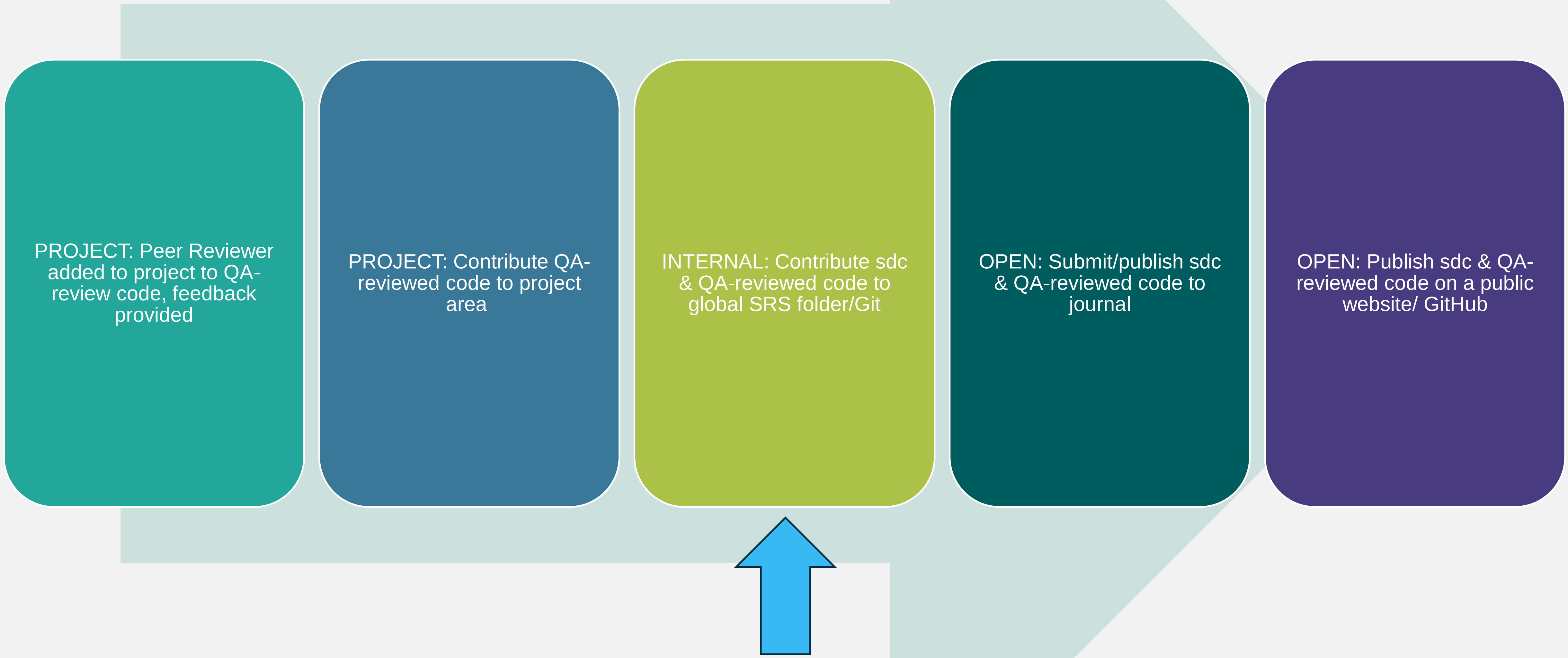
SRS check outputs to safeguard the confidentiality of data subjects

Output checking course is available to all support staff to ensure they are appropriately trained to manage outputs

Research serving the public good is published

- ● Safe People
- ● Safe Projects
- ● Safe Data
- ● Safe Setting
- ● Safe Outputs

Options for code sharing



PROJECT: Peer Reviewer added to project to QA-review code, feedback provided

PROJECT: Contribute QA-reviewed code to project area

INTERNAL: Contribute sdc & QA-reviewed code to global SRS folder/Git

OPEN: Submit/publish sdc & QA-reviewed code to journal

OPEN: Publish sdc & QA-reviewed code on a public website/ GitHub

Aim of pilot work

- Facilitate planning - code sharing group set up and manager role recruited
- Locate suitable use cases and start investigations/solutions

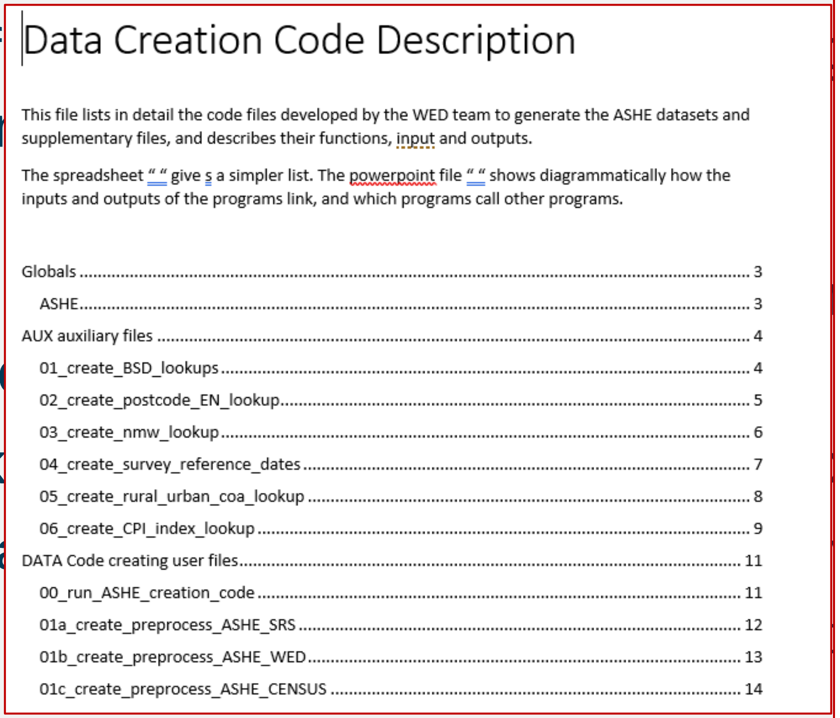
Explore and develop workable processes and protocols:

- ➢ 'As is' and proposed workflows
- ➢ Governance and administration, resourcing
- ➢ SRS and user policies
- ➢ User guidance and templates
- ➢ Capability building activity: webinars, 1-1 drop in sessions, blogs and case studies

- Roll out early adopter call and training sessions

Use case 1: Wealth and Employment Dynamics in Britain (WED) project

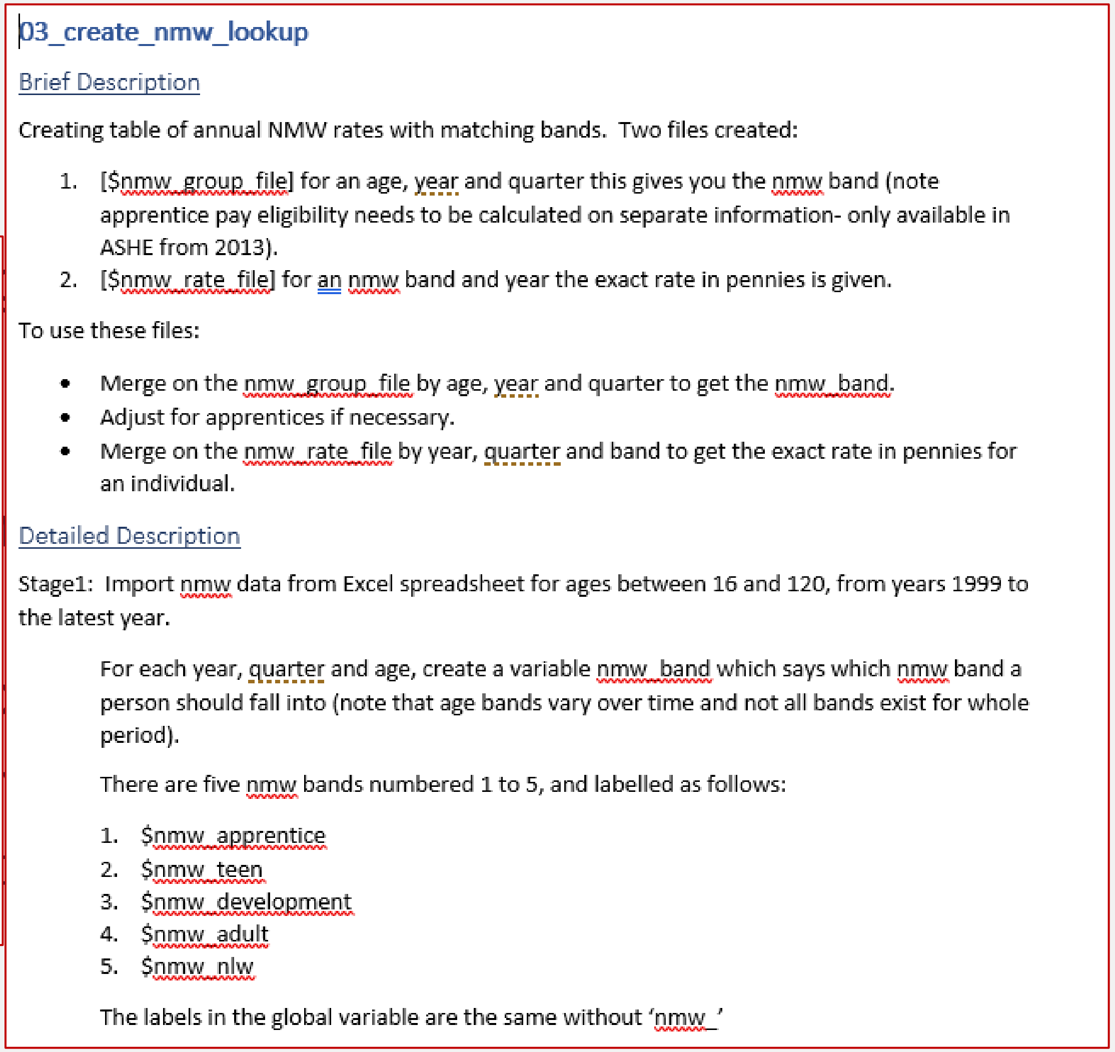
- Project aims to transform understandi
  Britain, f
  retiremer
- Involves
  (ASHE),
  occupati
- Project k
  - Data ma
  - Testing



Data Creation Code Description

This file lists in detail the code files developed by the WED team to generate the ASHE datasets and supplementary files, and describes their functions, input and outputs.

The spreadsheet “ “ give s a simpler list. The powerpoint file “ “ shows diagrammatically how the inputs and outputs of the programs link, and which programs call other programs.



03_create_nmw_lookup

Brief Description

Creating table of annual NMW rates with matching bands. Two files created:

1. [$nmw_group_file] for an age, year and quarter this gives you the nmw band (note apprentice pay eligibility needs to be calculated on separate information- only available in ASHE from 2013).
2. [$nmw_rate_file] for an nmw band and year the exact rate in pennies is given.

To use these files:

- Merge on the nmw_group_file by age, year and quarter to get the nmw_band.
- Adjust for apprentices if necessary.
- Merge on the nmw_rate_file by year, quarter and band to get the exact rate in pennies for an individual.

Detailed Description

Stage1: Import nmw data from Excel spreadsheet for ages between 16 and 120, from years 1999 to the latest year.

For each year, quarter and age, create a variable nmw_band which says which nmw band a person should fall into (note that age bands vary over time and not all bands exist for whole period).

There are five nmw bands numbered 1 to 5, and labelled as follows:

1. $nmw_apprentice
2. $nmw_teen
3. $nmw_development
4. $nmw_adult
5. $nmw_nlw

The labels in the global variable are the same without ‘nmw_’

Use case 2: Longitudinal Educational Outcomes dataset

- LEO is a de-identified, person level administrative dataset that brings together data on individual’s **education, employment, earnings data and benefits claims**
- Asset links data provided by **five separate government departments** via the SRS
- Dataset has the potential to provide transformative insight and evidence on the longer-term labour market outcomes /educational pathways of @**38 million English learners**
- 9 projects /50 researchers using data, including government users
- Early data manipulation work to create '**research-ready 'datasets/new variables**
- Some R code shared across ONS-led projects; target wider sharing in the SRS

Use case 3: Large scale Covid survey analysis

- April 2020 new survey launched from ONS, Universities, Public Health England: the Coronavirus Infection Survey (CIS), available in the SRS
- Interviews with each individual in a household, including nose and throat swabs (infection rate) and blood samples (antibodies)
- Large project with 80 researchers with **urgency and significant modelling asks**. Directly **informed government decision-making**
- Varied software use: R users preferred ONS Google Cloud Platform with less granular data
  - GCP uses GitHub to share code - ten repositories set up for each key analytical pipeline
  - SRS - initial poor management of code; later built basic version-controlled code using a master folder
- Review the code repositories for **lessons learned** for large multi-sector projects
- Work with data owner to review **publishing of analysts' code** in the SRS; distinguish added-value work from formal data documentation

Useful ONS resources

- Blog: https://www.gov.uk/government/news/coding-from-zero and https://intranet.ons.statistics.gov.uk/blog/coding-from-zero/
- Quality Assurance of Code for Analysis and Research
- Reproducible Analytical Pipelines (RAP) Champions
- Data Accelerator programme
- Reproducible Analysis — Coding for Economists (aeturrell.github.io)
- Tips for Better Coding — Coding for Economists (aeturrell.github.io)