Reproducible Analytical Pipelines Strategy

Building analysis as code

Martin Ralphs

Head of Analysis Standards and Pipelines

Methods and Quality Directorate

Office for National Statistics

Analytical pipeline

A process that produces an analytical product from data

Dataset

Vital statistics in the UK: births, deaths and marriages

Contact:
Anne Baker, Jon Darke, Rebecca Holley and Alex Howland

Release date:
24 February 2023

Next release:
January 2024

About this Dataset

Annual UK and constituent country figures for births, deaths, marriages, divorces, civil partnerships and civil partnership dissolutions.

Edition in this dataset

Current edition of this dataset

xlsx (349.6 KB)

Previous versions of this data are available.

Important notes and usage information

Main points from latest release

View all data related to population estimates

Contact details for this dataset

Anne Baker, Jon Darke, Rebecca Holley and Alex Howland
health.data@ons.gov.uk
+44 1329 444110

Methodology

User guide to divorce statistics

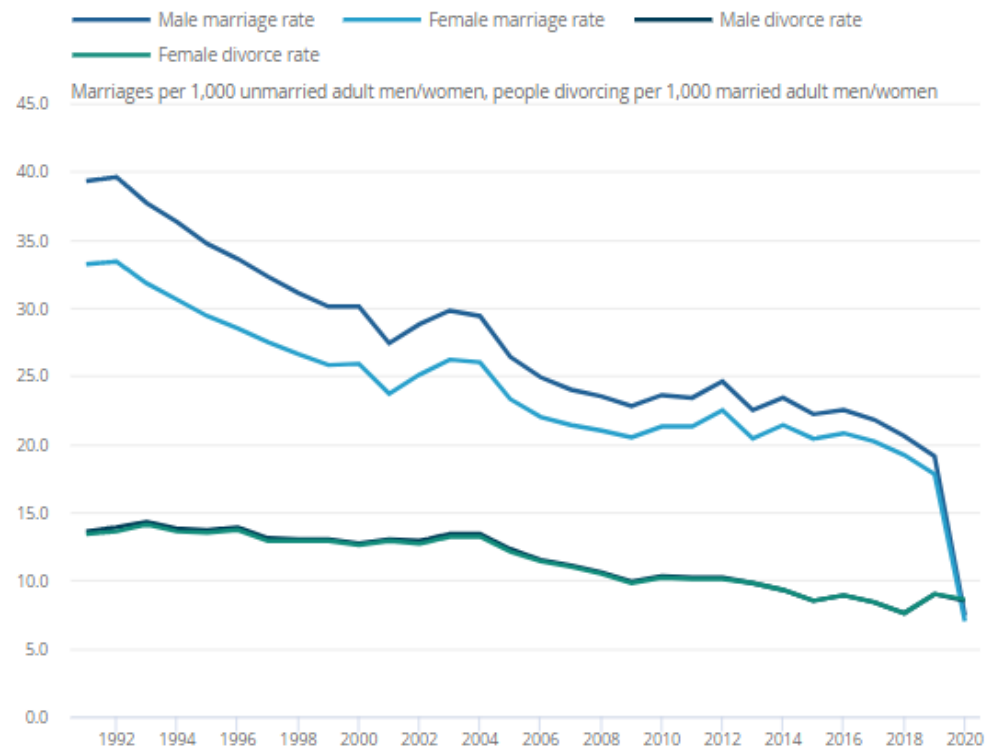
User guide to birth statistics

User guide to mortality statistics

User guide to marriage statistics

Figure 1: Marriage rates decreased by more than half in 2020 and were lower than divorce rates

Total marriage and divorce rates by sex, England and Wales, 1991 to 2020



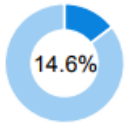
Source: Marriages in England and Wales from the Office for National Statistics

Notes:

1. Figures for 2014 onwards include opposite-sex and same-sex marriages, which have been possible in England and Wales from 29 March 2014.
2. The first divorces of same-sex couples took place in 2015.
3. These figures exclude civil partnership conversions.

DCMS Sectors Economic Estimates 2017 (provisional): Gross Value Added

DCMS contribution: £267.7bn



of UK GVA

increase since 2016

- In 2017, All DCMS Sectors contributed **£267.7bn** to the UK economy, accounting for **14.6% of UK GVA** (expressed in current prices).
- The GVA of DCMS Sectors has seen an increase of 3.4% since 2016 (£258.9bn in 2016) compared to 4.8% for the UK economy as a whole.
- The Digital Sector contributed £130.5bn to the UK economy in 2017, accounting for 7.1% of UK GVA. The contribution from this sector has increased by a third since 2010 (£98.2bn in 2010).
- The Creative Industries contributed £101.5bn to the UK economy in 2017, an increase of 53.1% since 2010 (£66.3bn).
- The Cultural Sector contributed £29.5bn to the UK economy in 2017, an increase of 38.5% since 2010 (£21.3bn).
- The Telecoms and Sports sectors saw increases of 31.6% and 40.0% respectively since 2010.
- The Gambling and Civil Society (non-market charities) sectors increased by 10.3% and 24.1% respectively since 2010.
- The Tourism Sector contributed £67.7bn to the UK economy in 2017, accounting for 3.7% of UK GVA.

This release provides estimates of the contribution of DCMS Sectors to the UK economy, measured by gross value added (GVA). Other economic measures, such as employment, trade and number of businesses are available in separate publications. These releases enable stakeholders to value the economic contribution of DCMS Sectors, which are not traditional National Account sectors, and to understand how current and future policy interventions can be most effective. The DCMS Sectors cover:

- Civil Society
- Creative Industries
- Cultural Sector
- Digital Sector
- Gambling
- Sport
- Telecoms
- Tourism

Note, the 2017 estimates are provisional and subject to change when National Accounts are published in 2019.

Responsible statistician:
Davita Patel
020 7211 2317

Statistical enquiries:
evidence@culture.gov.uk
@DCMSInsight

General enquiries:
enquiries@culture.gov.uk
0207 211 6200

Media enquiries:
020 7211 2210

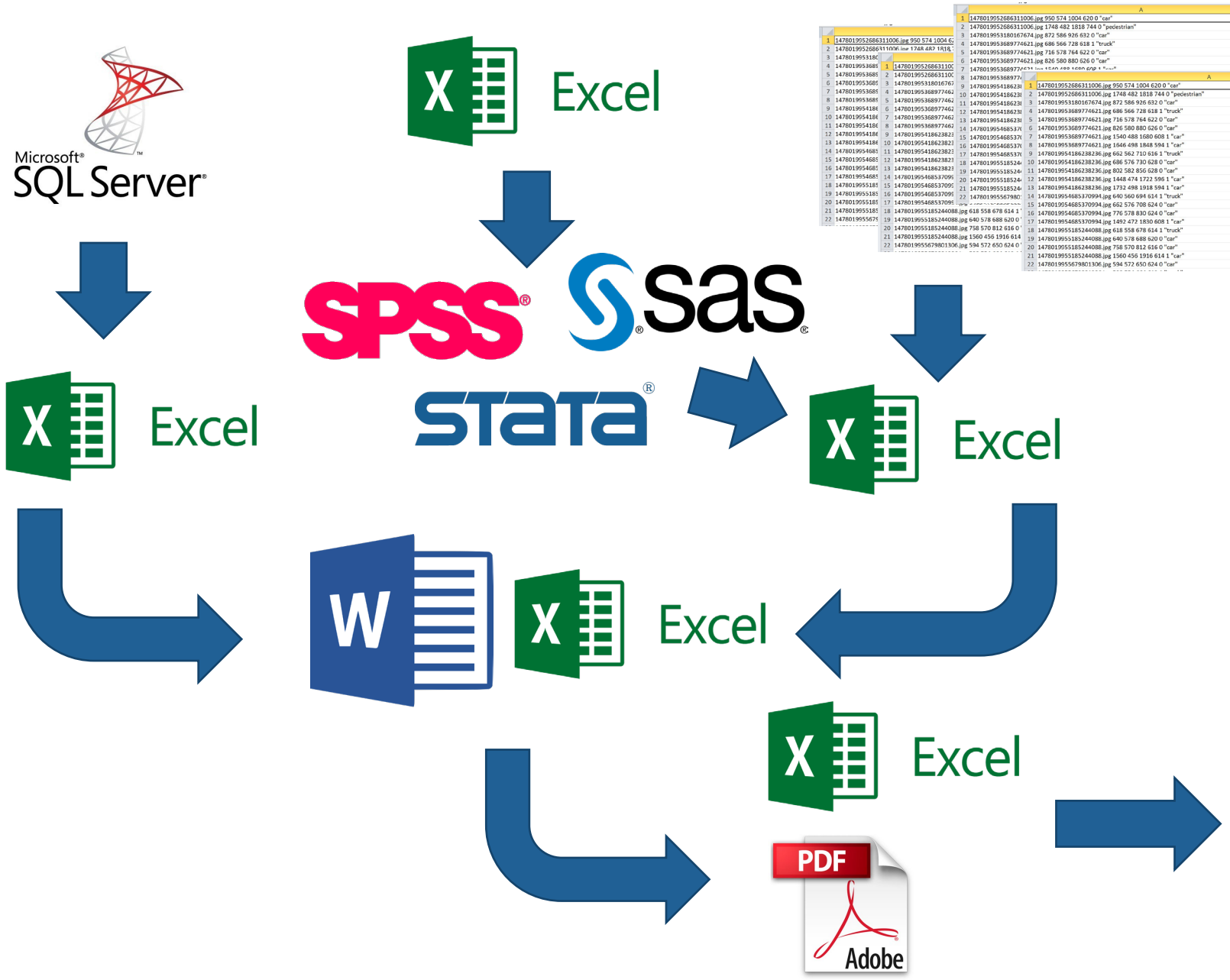
Contents

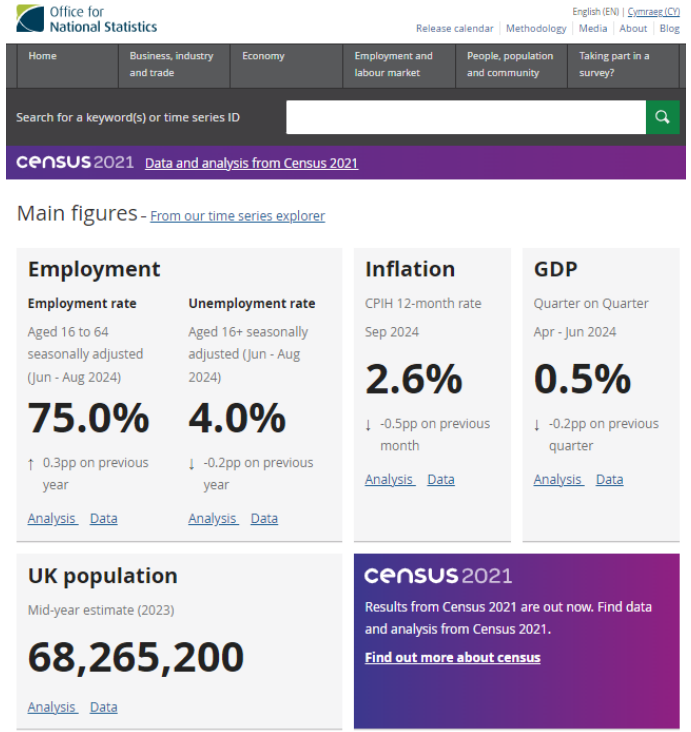
Table 2.1: GVA contribution (£bn, expressed in current prices) by DCMS Sectors: 2010 – 2017

Sector	2010(r)	2011(r)	2012(r)	2013(r)	2014(r)	2015(r)	2016(r)	2017(p)[1]	% change 2016 - 2017	% change 2010 - 2017	% of UK GVA 2017
Civil Society[2]	19.0	19.4	17.7	18.4	20.8	22.2	24.4	23.5	-3.7	24.1	1.3
Creative Industries	66.3	70.8	74.4	79.0	84.4	90.3	94.8	101.5	7.1	53.1	5.5
Cultural Sector	21.3	22.2	23.0	24.0	25.3	27.0	27.5	29.5	7.2	38.5	1.6
Digital Sector	98.2	103.9	106.1	111.4	113.1	115.0	121.5	130.5	7.3	32.9	7.1
Gambling	8.4	9.3	9.9	10.0	10.4	10.3	10.1	9.3	-8.2	10.3	0.5
Sport	7.0	7.4	7.9	7.5	7.8	8.7	9.3	9.8	5.3	40.0	0.5
Telecoms	24.8	25.5	26.0	28.1	30.0	30.4	31.4	32.6	3.6	31.6	1.8
Tourism[7]	49.2	53.9	57.3	59.0	60.4	68.0	68.3	67.7	-0.9	N/A	3.7
All DCMS Sectors (excl. Tourism)	**147.1**	**155.7**	**158.9**	**167.0**	**173.7**	**183.5**	**190.7**	**200.1**	**4.9**	**36.0**	**10.9**
All DCMS Sectors[7]	**196.3**	**209.6**	**216.2**	**226.0**	**234.2**	**251.5**	**258.9**	**267.7**	**3.4**	**N/A**	**14.6**
UK	*1,429.6*	*1,468.3*	*1,514.9*	*1,573.2*	*1,646.0*	*1,692.0*	*1,756.0*	*1,839.9*	*4.8*	*28.7*	*100.0*

Notes

1. 2017 GVA is based on the output measure of GVA to allow consistency with the sector measures for 2017. This is aligned to average GVA up to and including 2016 (last Supply Use balanced year) but then uses growth in the output measure as a proxy for GVA beyond that. The 2017 GVA figure will be revised next year, once the Supply Use tables have been balanced. This approach is different for Civil Society where the average proportion of the UK economy that is attributed to NPISH for 2010 to 2016 is assumed to be the same for 2017. This assumption seems reliable given the proportion does not vary much (from 1.2% to 1.4% over these years).
2. The Civil Society figure covers non-market charities in the NPISH (non-profit institutions serving households) sector. It does not include market provider charities who have passed the market test and therefore sit in the corporate sector (these data are not currently measured by ONS on a National Accounts basis), mutuals, social enterprises or community interest companies. Therefore, this is an underestimate for the sector.
3. DCMS Sector total is lower than the sum of individual DCMS Sectors because of overlaps between sectors.
4. p = provisional
5. r = revised. These are planned revisions and part of the annual adjustment and balancing process of National Accounts.
Text in red show where the data have been revised due to the balancing of Supply and Use tables or revisions of the NPISH data (affecting Civil Society Sector).
6. Data are in current prices (i.e. have not been adjusted for inflation).
7. Estimates for Tourism are based on a different methodology to all other sectors, as they are taken from the Tourism Satellite Account. Several methodology improvements were made for the 2016 Tourism data, which resulted in the 2015 data being revised. In 2016, several improvements were made to the Great Britain Day Visits Survey (GBDVS). More information on these changes can be found in Chapter 3 of the methodology note. As a result of these changes, a 15% increase was observed in the levels of visits reported by respondents. The 2015 data have been revised in line with the increased level of reporting of day visits. This change has not yet been implemented in the data prior to 2015. ONS plan to implement these changes in 2019 and therefore, caution should be taken when comparing data from 2015 onwards with previous years.


Microsoft SQL Server
Excel
SPSS
sas
STATA
Excel
Excel
Excel
Excel
PDF
Adobe


Office for National Statistics
English (EN) | Cymraeg (CY)
Release calendar | Methodology | Media | About | Blog
Home
Business, industry and trade
Economy
Employment and labour market
People, population and community
Taking part in a survey?
Search for a keyword(s) or time series ID
census 2021 Data and analysis from Census 2021
Main figures - From our time series explorer
Employment
Employment rate
Aged 16 to 64 seasonally adjusted (Jun - Aug 2024)
75.0%
↑ 0.3pp on previous year
Analysis Data
Unemployment rate
Aged 16+ seasonally adjusted (Jun - Aug 2024)
4.0%
↓ -0.2pp on previous year
Analysis Data
Inflation
CPIH 12-month rate
Sep 2024
2.6%
↓ -0.5pp on previous month
Analysis Data
GDP
Quarter on Quarter
Apr - Jun 2024
0.5%
↓ -0.2pp on previous quarter
Analysis Data
UK population
Mid-year estimate (2023)
68,265,200
Analysis Data
census 2021
Results from Census 2021 are out now. Find data and analysis from Census 2021.
Find out more about census

Manual pipelines carry quality risks

Outputs take a long time to produce and quality assure and a long time to reproduce

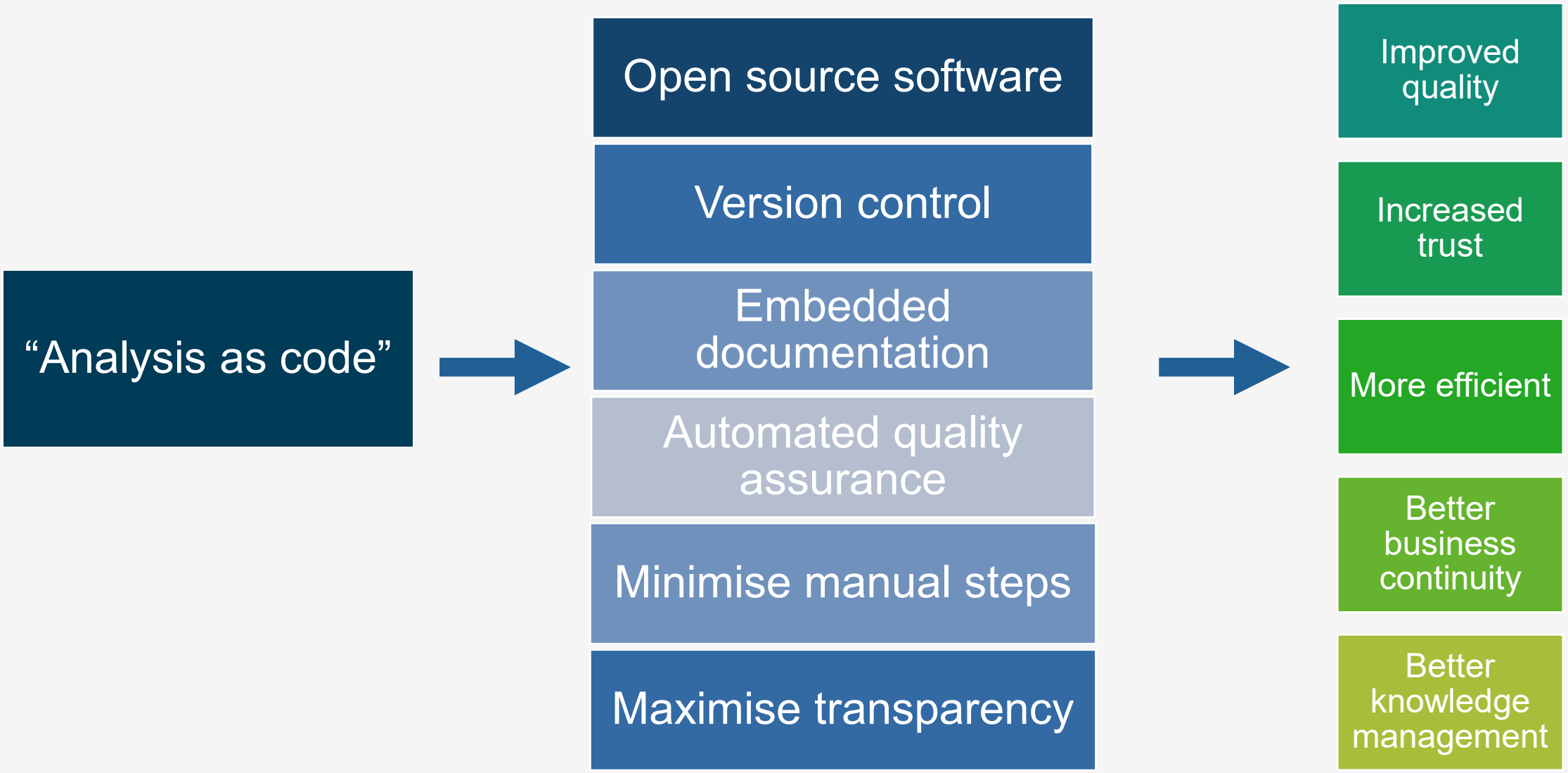
Source data and outputs are not connected, except through manual steps

Processes are manual, hard to follow and tedious, increasing the risk of mistakes

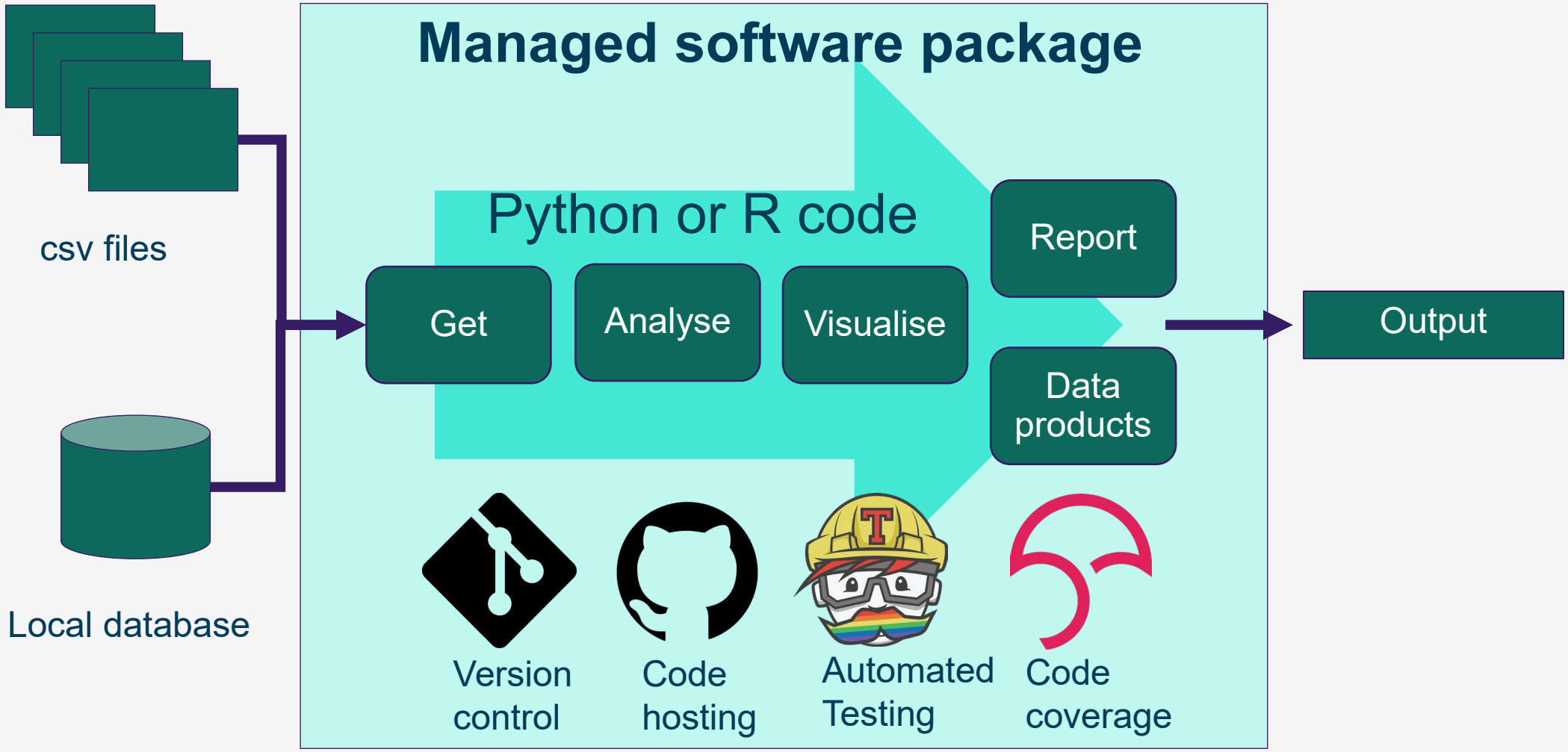
“Copy and paste” and repetitive manual steps are error-prone

It is hard to track changes without a lot of manual effort

Reproducible Analytical Pipelines



A reproducible analytical pipeline



What does this look like in practice?

- Bulletins built automatically by code, along with charts, tables, datasets and supporting documents
- Automatic QA reports, logs and validation
- Datasets formatted, validated and built automatically
- Code hosted remotely, with version control so everybody working on it knows **who did what**, **when** and **why**

```{r Diagnostic graph, fig.width = 10, fig.height= 5, dpi=200, dev = "win.metafile"}
#GRAPH
graph <- ggplot2::ggplot(data) +
  ggplot2::geom_bar(
    ggplot2::aes(x = date, y = daily_deaths),
    stat = "identity",
    colour = "#293271",
    fill = "#293271"
  ) +
  ggplot2::geom_line(ggplot2::aes(x = date, y = rolling_average),
                     colour = "#d8852b",
                     size = 2.5) +
  ggplot2::theme_bw() +
  ggplot2::labs(y = "Deaths by date of death",
                x = "")
graph
```

Check data
A series of checks on the data

Column Names

```{r Check the column names}

if(length(covpress::check_column_names_correct(data, expected_column_names)) > 0) {
  cat(
    paste0(
      "Column names are not as expected. Standard column names are:\n",
      expected_column_names,
      " but there are differences here:\n"
    )
  )
  missing <-
    covpress::check_column_names_correct(data, expected_column_names)
} else {
  cat("Column names are as expected and will not be shown here.")
}
```

Rows or observations

There are `r nrow(data)` rows today. There were `r nrow(data_minus)` rows in the last release. Today's data should have at least `r additional_rows_expected` additional rows.

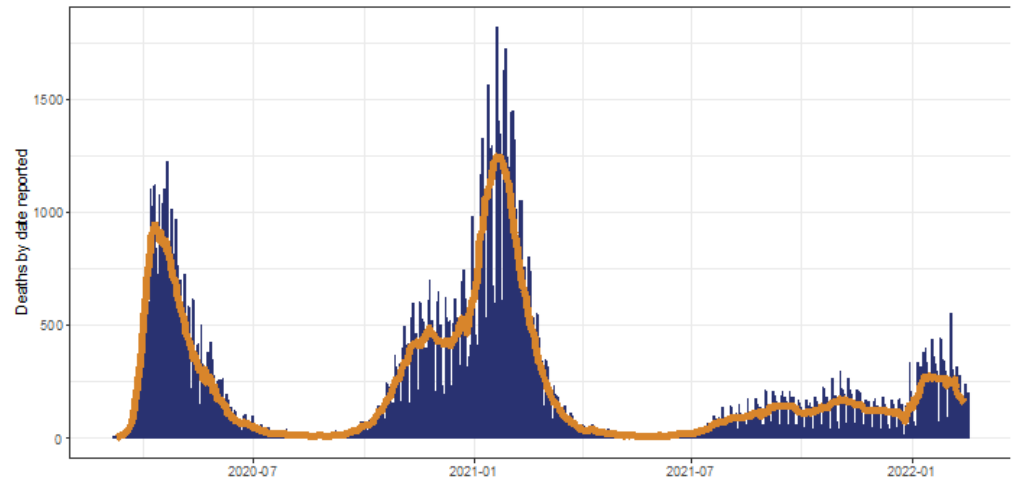
Missing values

Missing values are displayed below. The generated columns, *Rolling Average* and *Rolling Sum* are ignored for the missing value check.

```{r Check for missing values, echo=FALSE, comment=NA}
# Check columns selected are correct

if (nrow(data[!complete.cases(data[3:5]), ]) > 0) {
  covpress::format_column_names(data)

  table <-  flextable(data[!complete.cases(data[3:5]), ])
  table
```



Check data

A series of checks on the data

Column Names

Column names are as expected and will not be shown here.

Rows or observations

There are 713 rows today. There were 712 rows in the last release. We would expect to see at least 1 more row(s) in today's release.

Missing values

Missing values are displayed below for numeric columns in the data. The generated columns, *Rolling Average* and *Rolling Sum* are ignored for the missing value check.

area_name	date	daily_deaths	cumulative_deaths	daily_death_change	rolling_average	rolling_sum
united_kingdom	2020-03-06	1	1			

RAP benefits: efficiencies

- RAPs reduce processing time and the resource needed to produce statistical outputs
- Typical FTE savings of from 50-95% of analyst time
- Very labour-intensive manual processes see the biggest benefits, especially if they are run regularly
- RAPs free up analyst time to do more analysis and less tedious, risk-prone manual work

RAP benefits: significant quality improvement

- Processes are well documented, so easier to understand
- Processes are faster to maintain and fix
- Processes are easier to pick up, so more resilient
- We can re-use modular components
- RAPs have helped us move our work to cloud-based environments more quickly
- RAP can be applied even for very high-pressure work

BUT RAPs require maintenance and updating!

Strategic enablers for RAP

- “Open source by default” policy for UK government
- Analysis Function RAP Strategy to deliver analysis using RAP by default. Three strands focus on tools, capability and culture
- Active communities of practice like the RAP champions' network and use of peer review
- Tools, guidance, standards and policies promote RAP practices
- Shared examples of “what good looks like”
- Consultancy and mentoring support for analyst teams
- RAP learning pathway to build capability

Our approach to building capability

- Start small and grow incrementally
- Develop early examples to demonstrate value and impact
- Teams learn by doing not by sitting through courses
- We use “just-in-time learning” so training is used immediately
- Teams learn together, through paired development and mentoring support
- Use good practice from the beginning
  - ✓ Version control and code hosting
  - ✓ Coding standards (like PEP8 for Python or tidverse for R)
  - ✓ Built-in testing
  - ✓ Comprehensive documentation
  - ✓ Packaged, modular code

Implementation models for RAP

We use different approaches to meet different needs:

- Hub and spoke model to build and embed capability via central consultancy and support function which sets standards and guidance
- Local business area teams to build local RAPs
- Expert, dedicated teams to support major projects
- Crisis / surge function for rapid response

The main challenges

- Skills retention, in teams and organisations
- Capability – getting to a critical mass
  - Of coding competency
  - Of managers who can assure RAPs
- Developing a culture that promotes “analysis as code” as the standard way to build analysis
- A technology stack that enables RAP practices
- Risk aversion to developing in the open

RAP works well when

Senior managers give commitment and advocacy

Team members are committed to the work

Teams have enough time to contribute

There is a base level of technical understanding

The right tools are in the right place

There is a plan to move to business as usual with resource to maintain and update pipelines

There is a shared view of what good looks like

There is a supportive community of practice

RAP is harder to do when

RAP is not seen as a positive culture change

There is limited access to open-source tools

Coding capability is limited

Not enough time is set aside to do RAP

There is no plan for sustainability

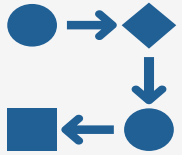
Poorly written and managed analysis code is as risky as manual analysis!

The code is hard to understand	So it's hard to use and hard to assure
The code is repetitive	Likely to contain mistakes, hard to change and adapt
Manual version control or no version control at all	We don't know who changed what, when or why We can't revert to earlier versions We can't keep track of changes
The code is not tested	We can't be sure the code performs as expected
Manual intervention during the run	Manual steps lead to human error and increase risk

RAP Guidance

Quality Assurance of Code for Analysis and Research (QACAR)

Sets out good practices for writing reproducible, transparent and resilient code



Minimum Viable Product for RAP

Minimum application of software engineering practice to analysis

Reflects feedback from the RAP Champions

ONS version is more stringent

Code QA Checklists

Reflects that quality assurance of code should be proportionate (AQuA book)

Reiterates the content of the RAP minimum viable product