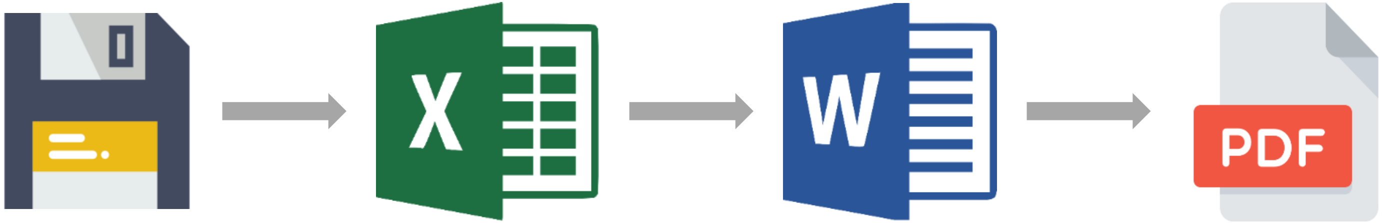
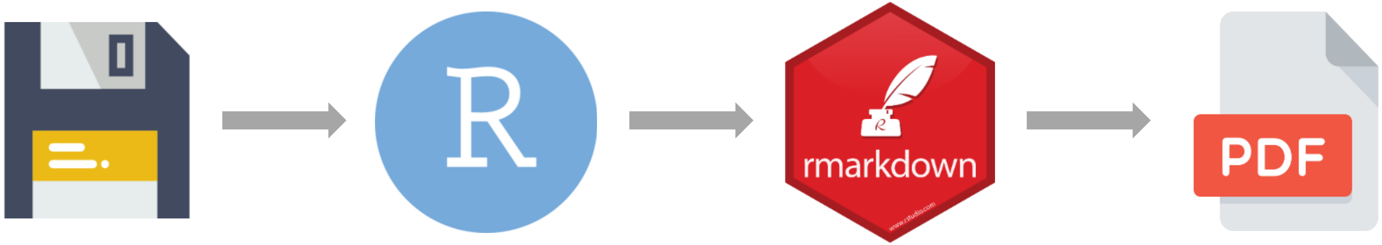
Institute
and Faculty
of Actuaries

Reproducible Data Science: What Can We Learn From Other Professions?

Philip Darke FIA
Dr Matthew Forshaw

26 September 2019


X
W
PDF

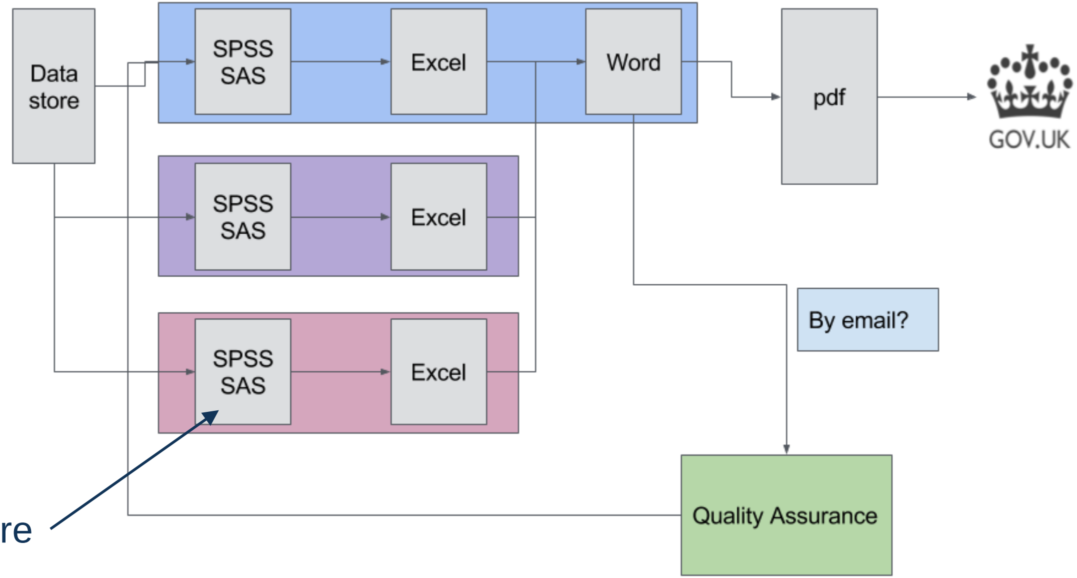

R
rmarkdown
www.rstudio.com
PDF

Institute
and Faculty
of Actuaries

Case study

Official statistics in UK Government

Official statistics in UK Government



In-house software

https://ukgovdatascience.github.io/rap_companion/why.html#the-current-statistics-production-process

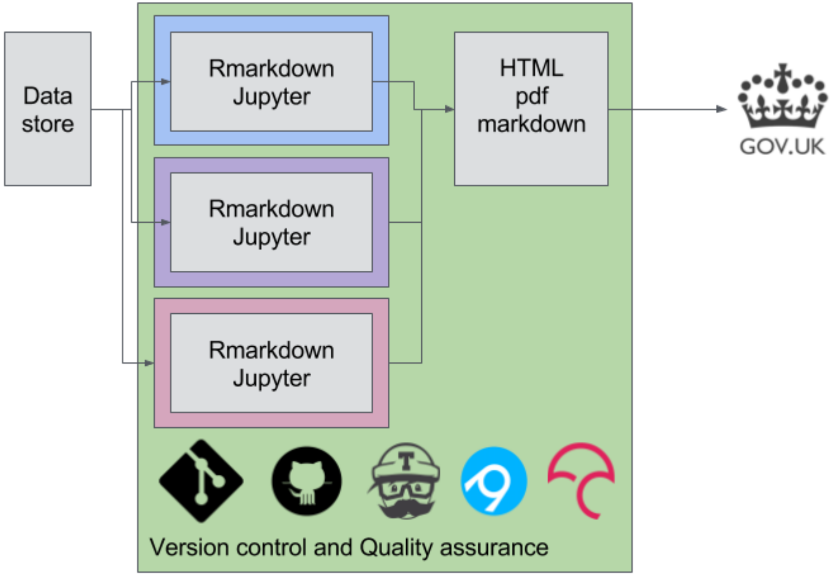
Institute and Faculty of Actuaries

Manual processes add risk

- Errors in spreadsheets are common
- Manual processes risk introducing human error
- Checking and peer review are not embedded in the process
- Challenging to reproduce previous work

Institute
and Faculty
of Actuaries

Reproducible analytical pipelines



https://ukgovdatascience.github.io/rap_companion/why.html#desired-reproducible-analytical-pipeline

Institute and Faculty of Actuaries

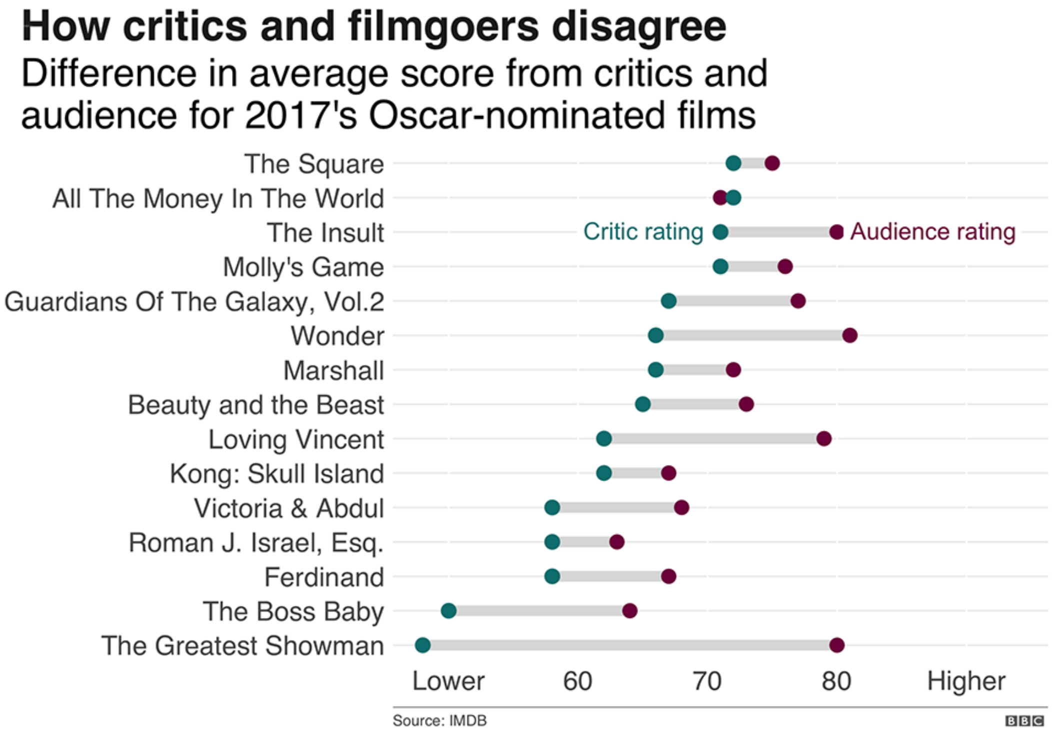
> “The potential **time savings** for analysts are enormous, freeing them up to focus on the interpretation of the results. The other huge benefit comes from building a process that is fully **transparent, auditable and verifiable** – reducing risk and improving quality.”

Matt Upson and Mat Gregory, Government Digital Service

https://dataingovernment.blog.gov.uk/2017/11/27/transforming-the-process-of-producing-official-statistics/

Institute
and Faculty
of Actuaries

BBC News website graphics



How critics and filmgoers disagree

Difference in average score from critics and audience for 2017's Oscar-nominated films

Source: IMDB

https://www.bbc.co.uk/news/entertainment-arts-43146027 (edited to fit)

Institute and Faculty of Actuaries

> [This approach] **saves a huge amount of time and effort**, in particular when working with data that needs updating regularly, with **reproducibility** a key requirement of our workflow. In short, it was a game changer…

BBC Visual and Data Journalism team

https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535

Institute
and Faculty
of Actuaries

Reproducible work

Reproducibility is the process of making code and data available so that others can easily replicate, verify and build on your analysis

Building blocks of a reproducible workflow

- Data
- Analytic code and automated checks
- Documentation
- Computational environment
- Packaged in a standard way

Why is this important for actuaries?

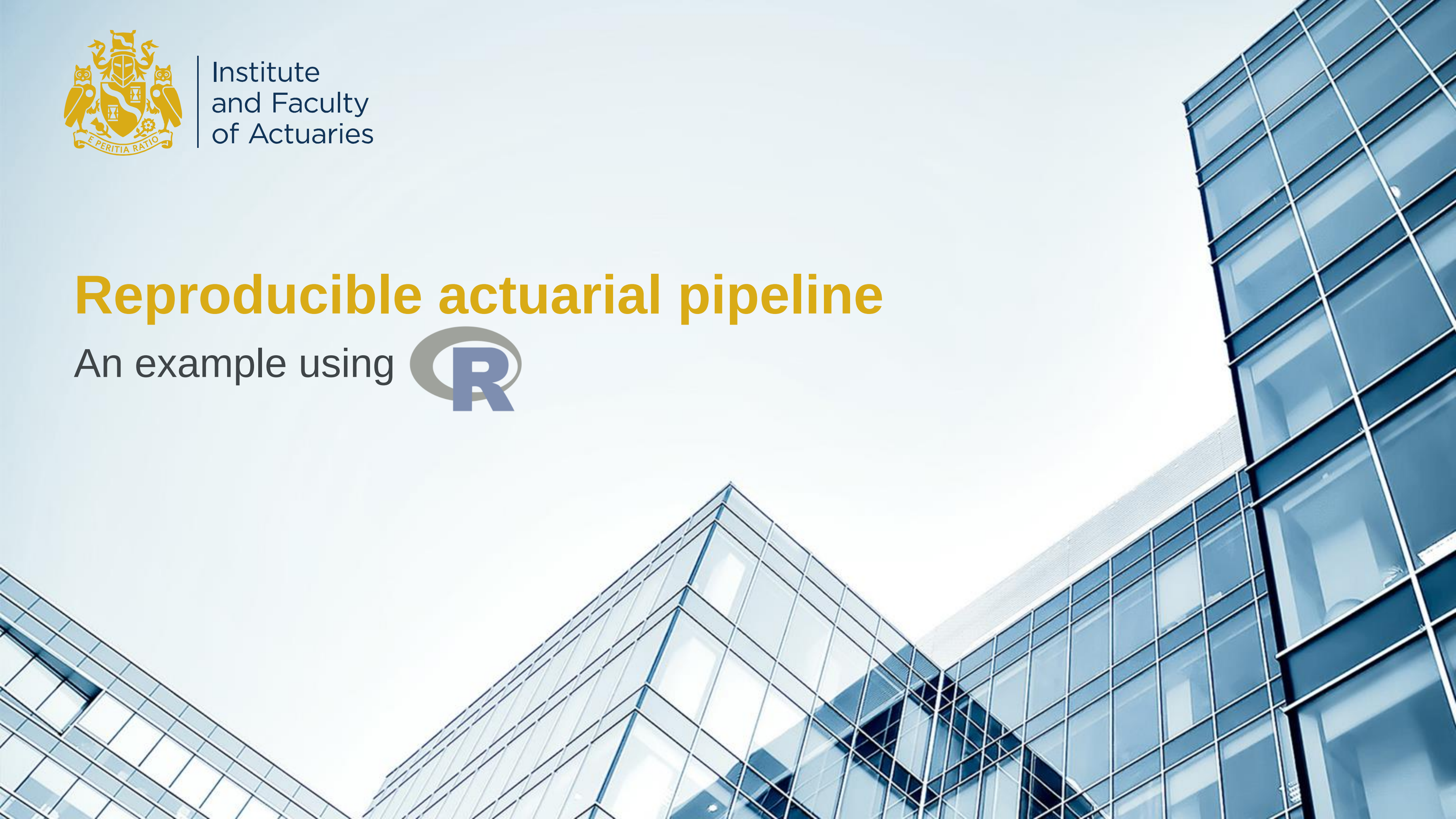
- Enables more efficient working
- Allows analysts to focus on the bigger picture
- Easier collaboration
- Helps meet compliance requirements – internal and TAS
- A step towards automation

Institute
and Faculty
of Actuaries

Institute
and Faculty
of Actuaries

Reproducible actuarial pipeline

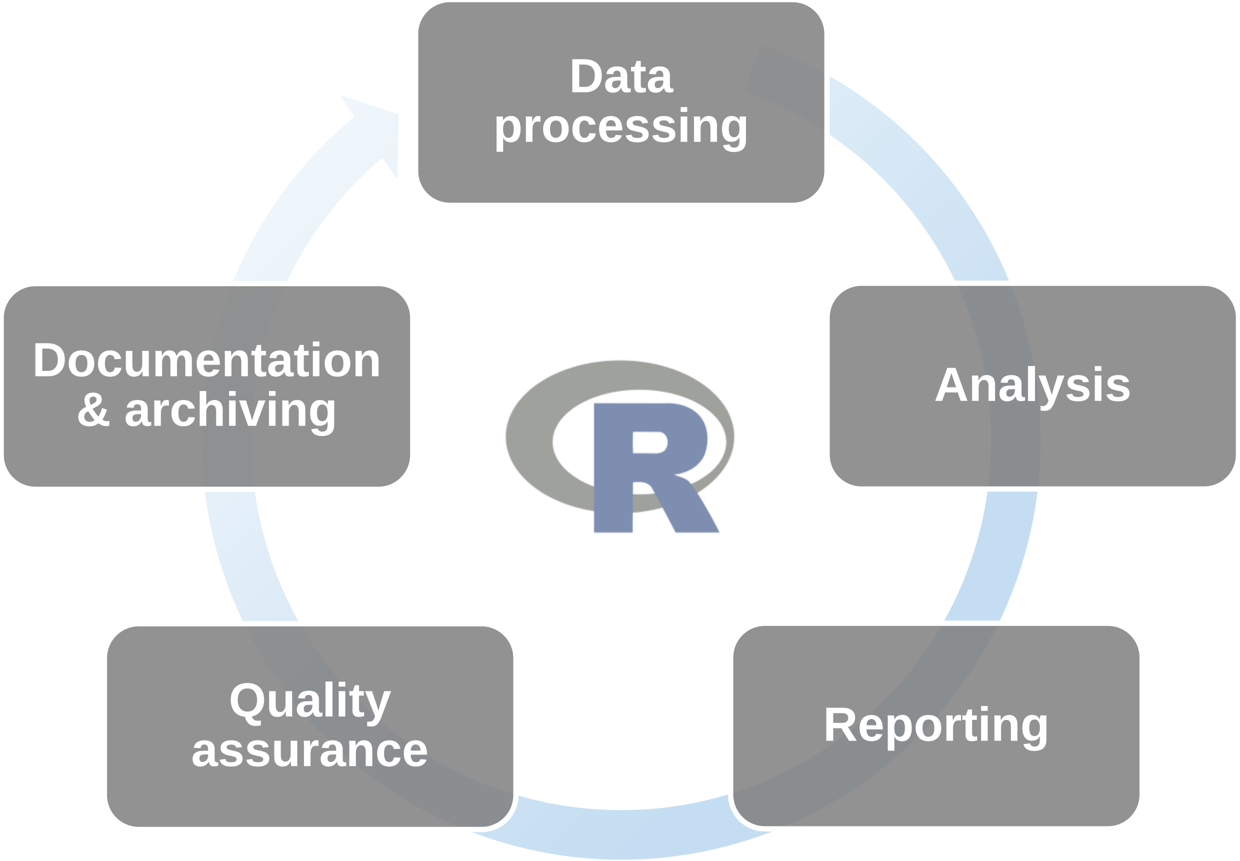
An example using R

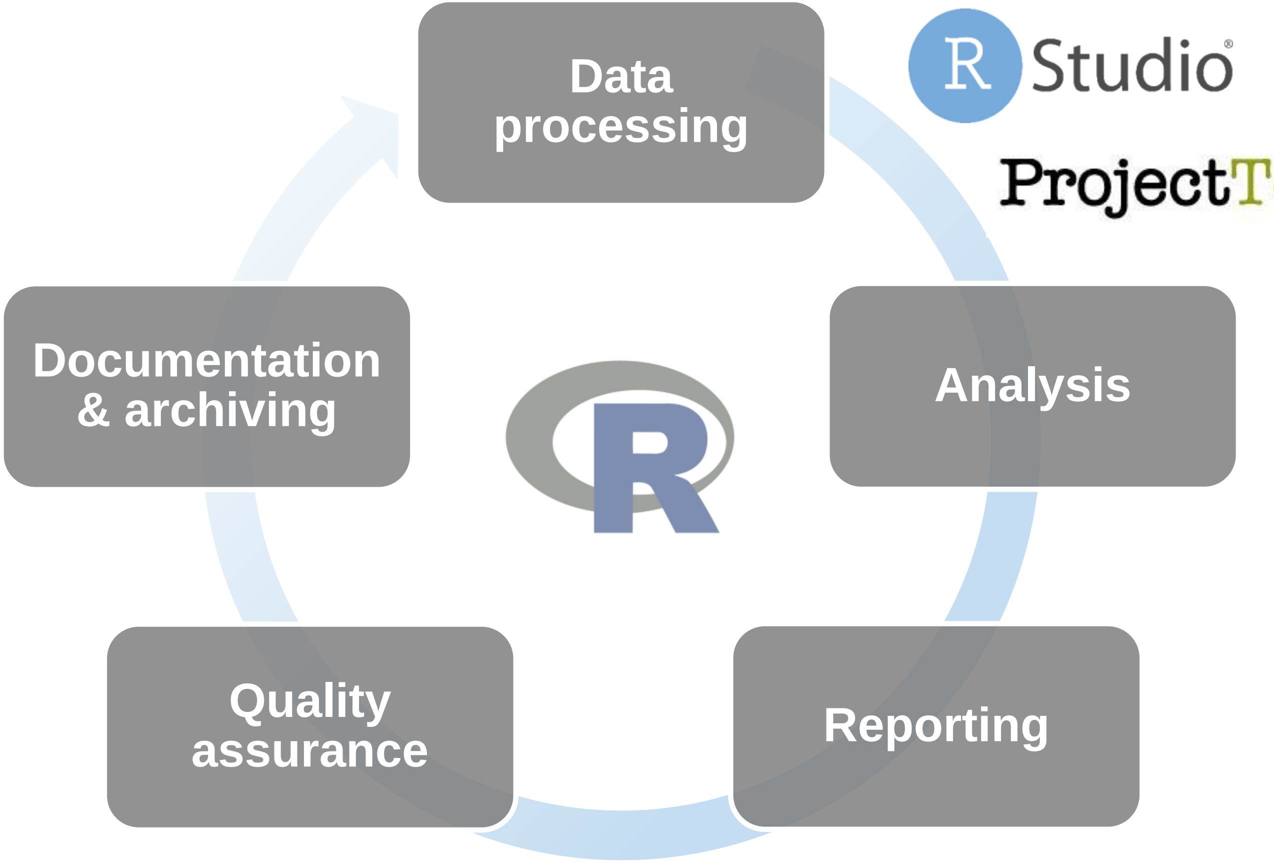


Why R?

- Stable, up-to-date and free
- Open source with an active support community
- Well suited to building reproducible pipelines and reporting
- Next generation of actuaries will learn R under the 2019 curriculum

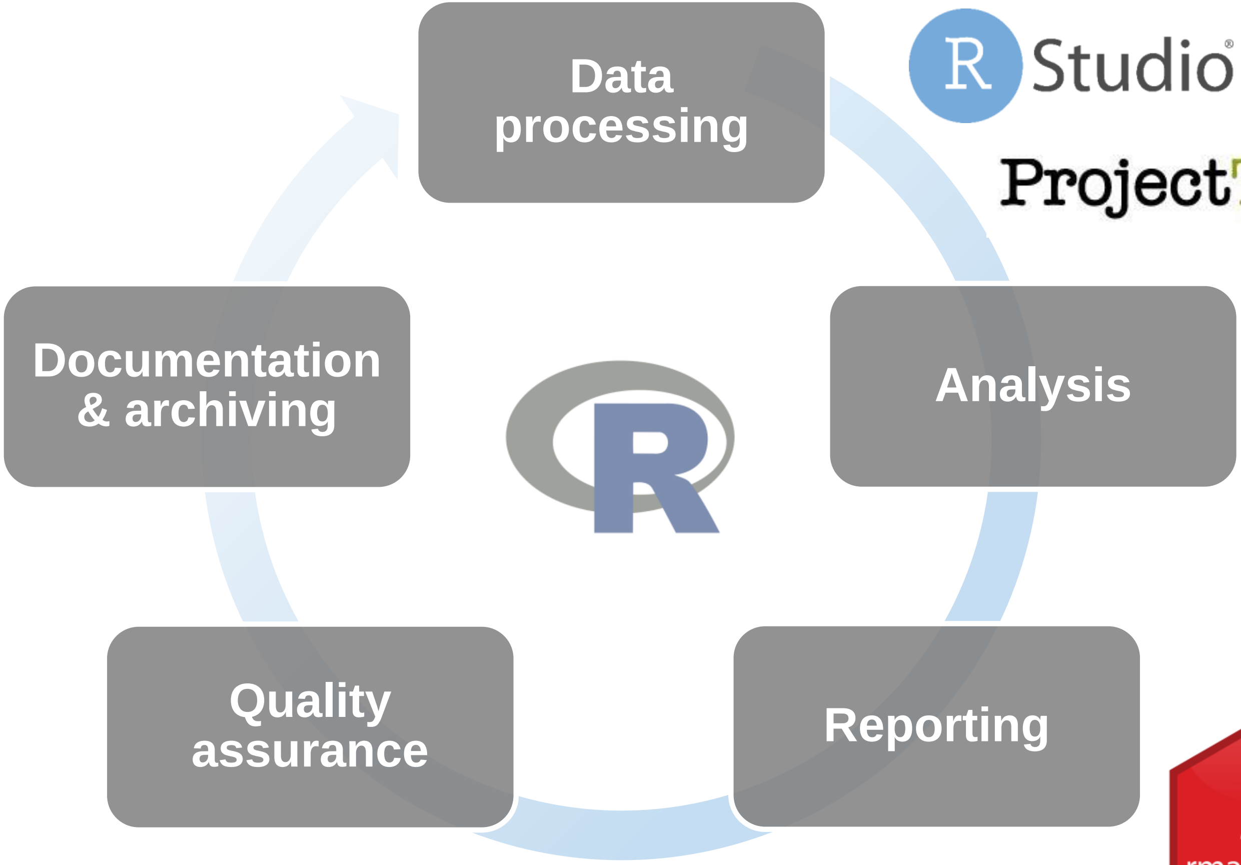
Institute
and Faculty
of Actuaries


Data processing
Analysis
Reporting
Quality assurance
Documentation & archiving



RStudio

ProjectTemplate



Data processing

Analysis

Reporting

Quality assurance

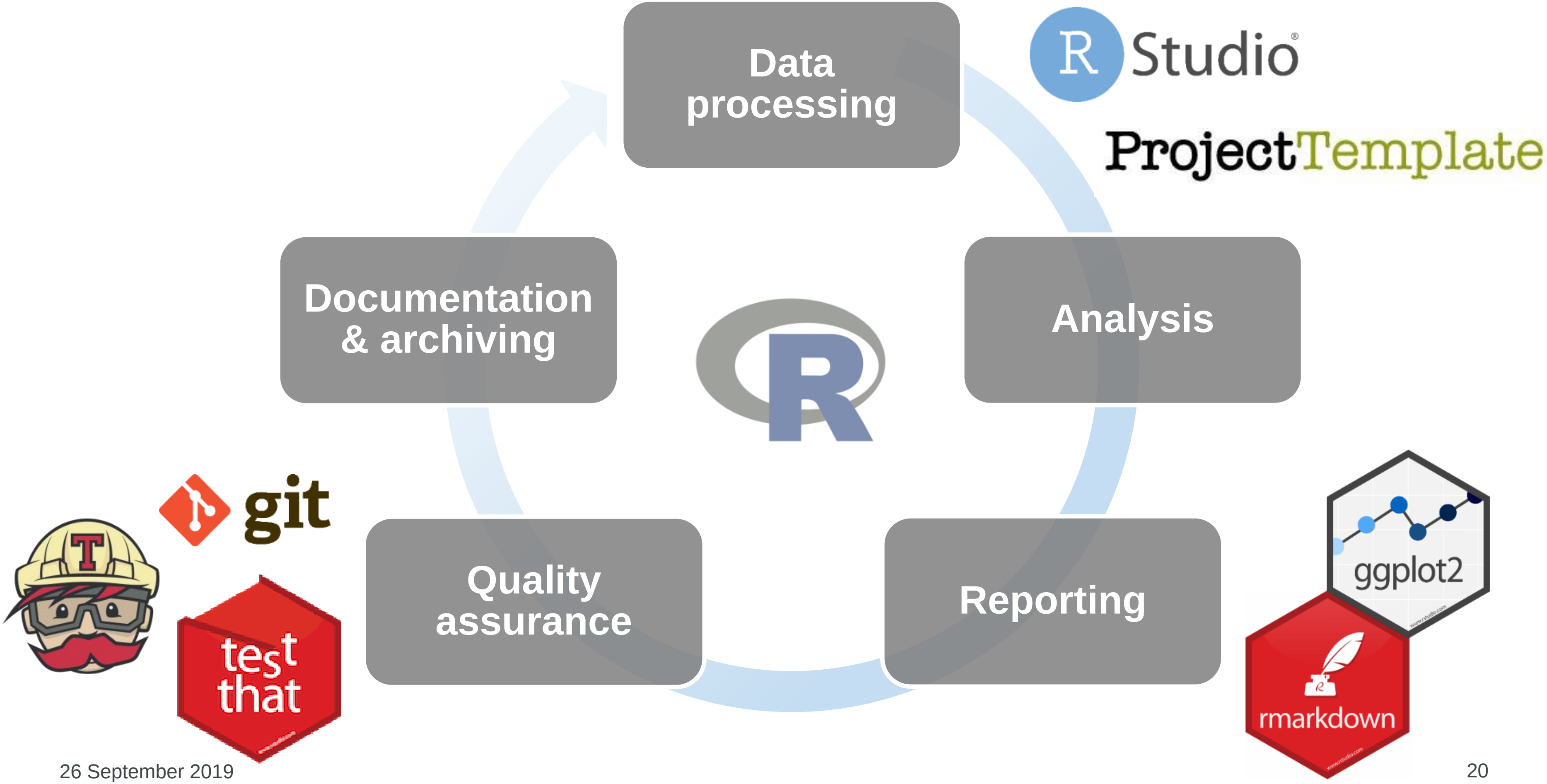
Documentation & archiving

RStudio

ProjectTemplate

ggplot2

rmarkdown



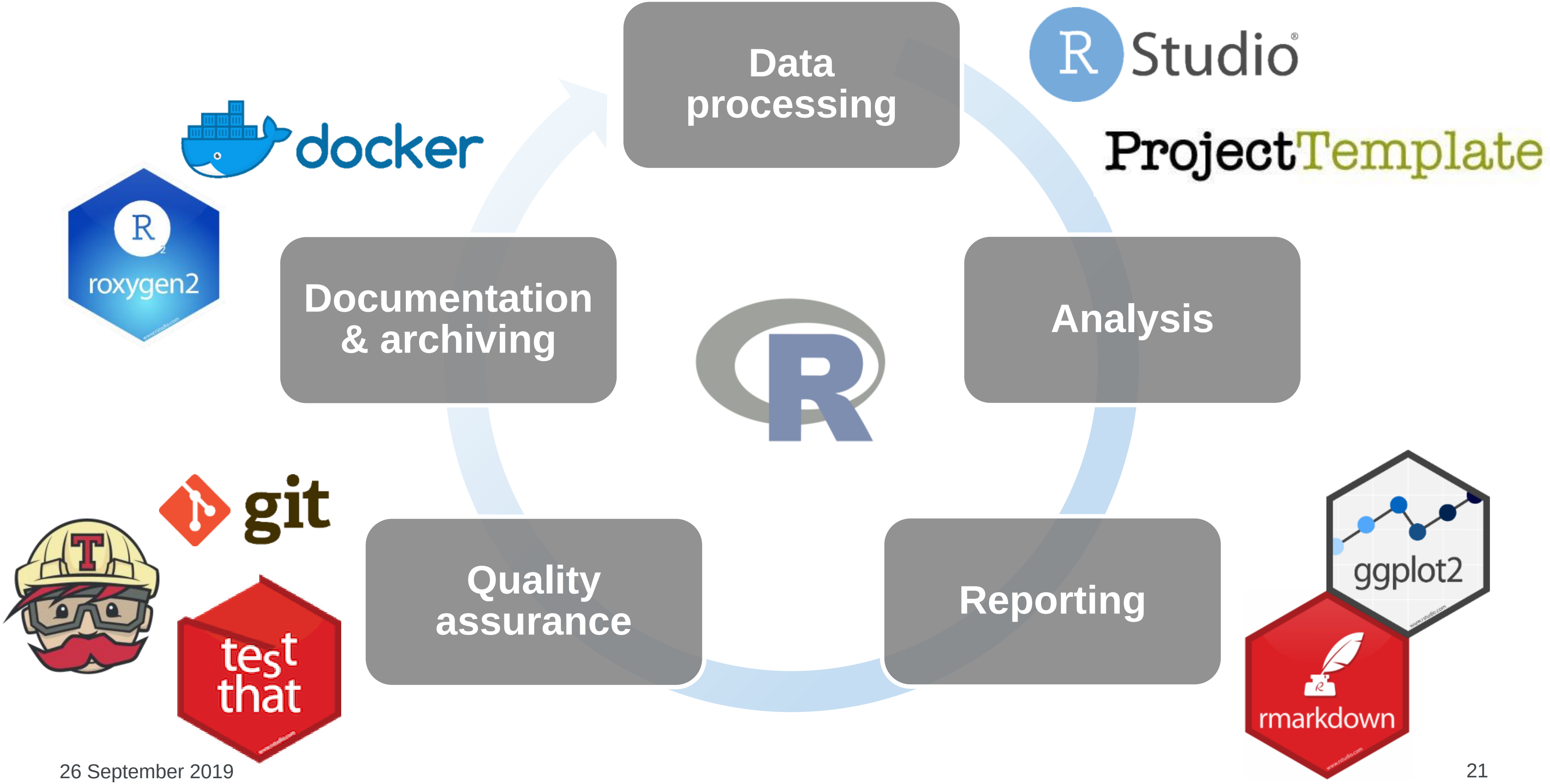
Data processing

Analysis

Reporting

Quality assurance

Documentation & archiving



Data processing

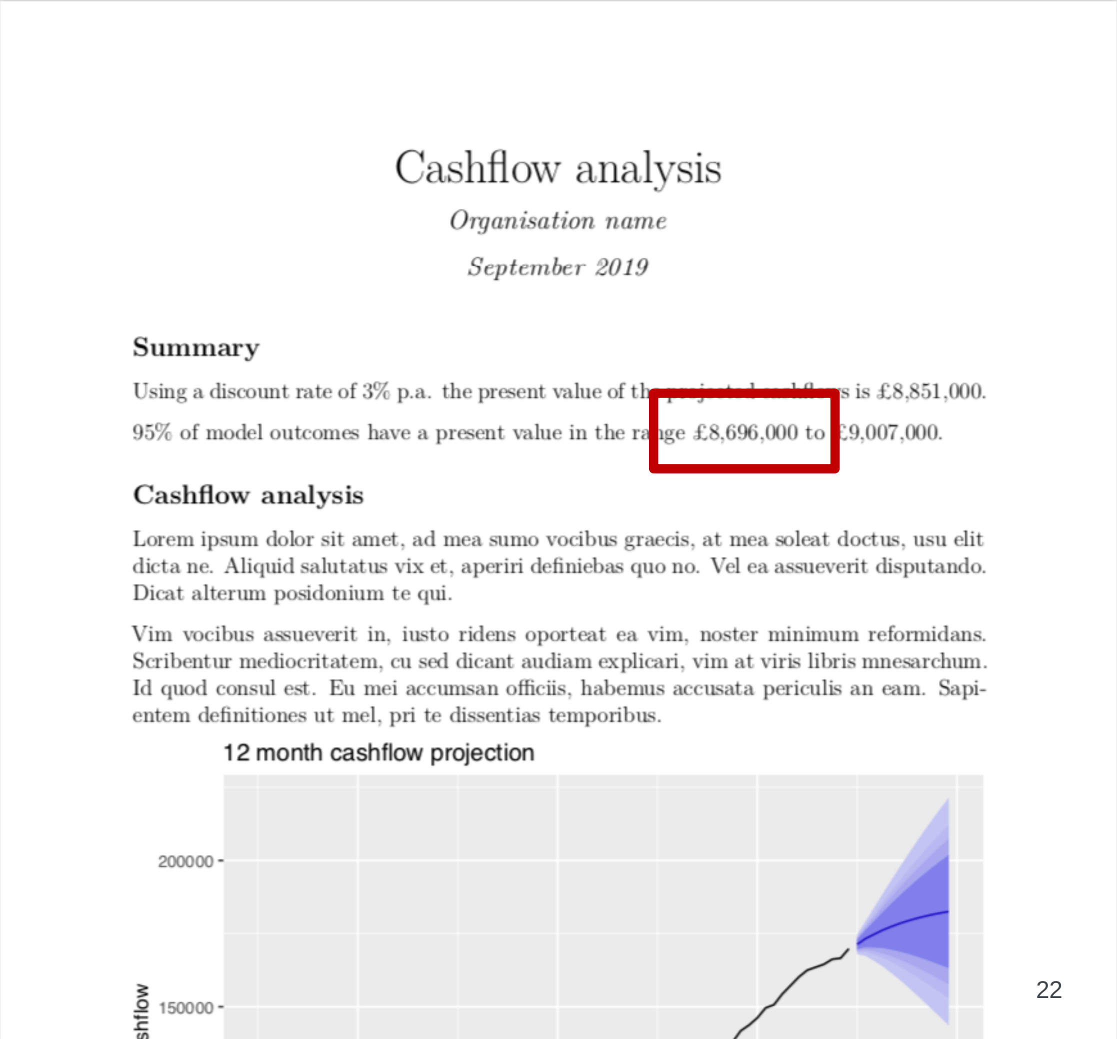
Analysis

Reporting

Quality assurance

Documentation & archiving

> “Where does that figure come from?”



Cashflow analysis

Organisation name

September 2019

Summary

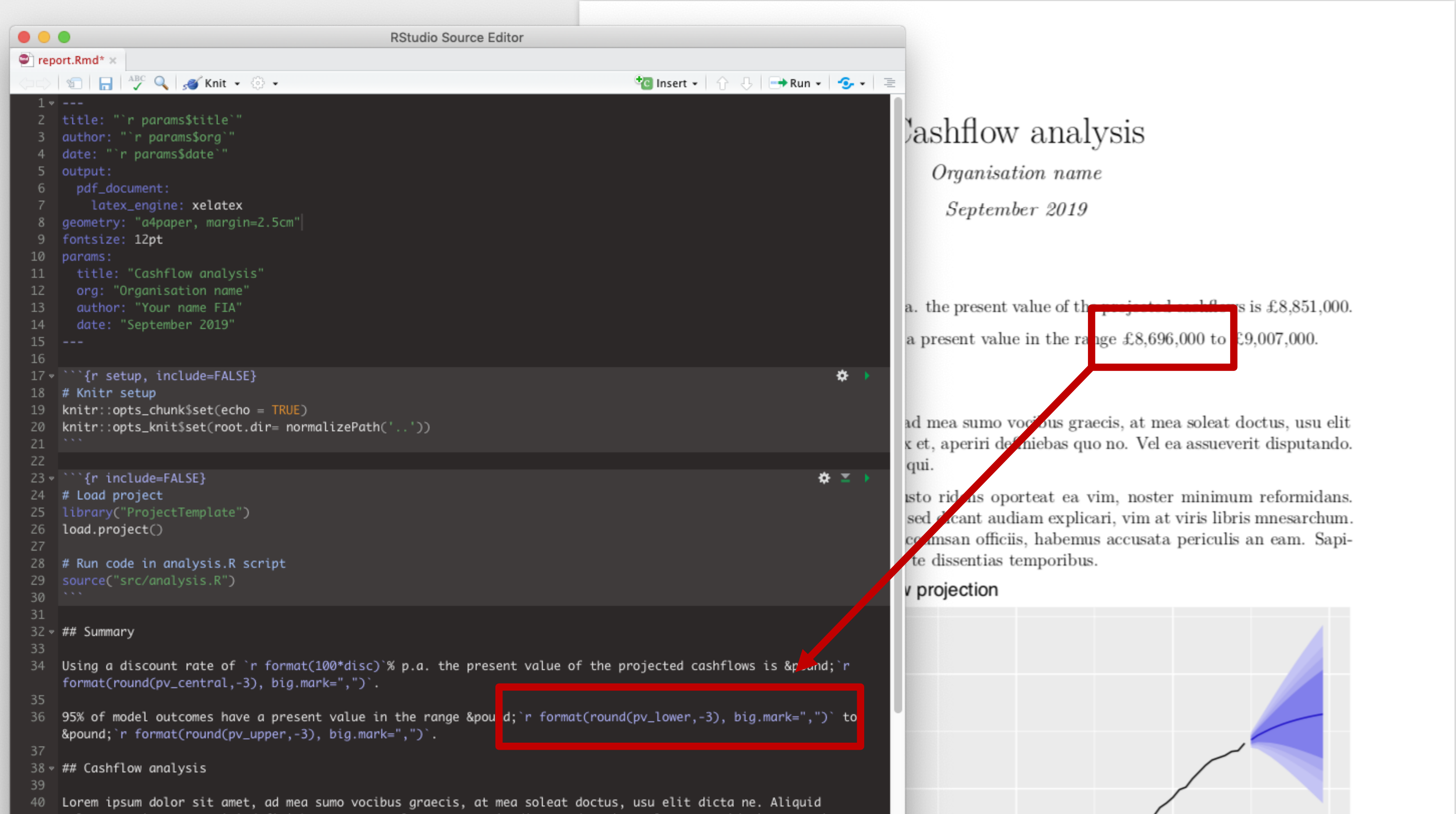
Using a discount rate of 3% p.a. the present value of the projected cashflows is £8,851,000.

95% of model outcomes have a present value in the range £8,696,000 to £9,007,000.

Cashflow analysis

Lorem ipsum dolor sit amet, ad mea sumo vocibus graecis, at mea soleat doctus, usu elit dicta ne. Aliquid salutatus vix et, aperiri definiebas quo no. Vel ea assueverit disputando. Dicat alterum posidonium te qui.

Vim vocibus assueverit in, iusto ridens oporteat ea vim, noster minimum reformidans. Scribentur mediocritatem, cu sed dicant audiam explicari, vim at viris libris mnesarchum. Id quod consul est. Eu mei accumsan officiis, habemus accusata periculis an eam. Sapientem definitiones ut mel, pri te dissentias temporibus.



~~~
---
title: "`r params$title`"
author: "`r params$org`"
date: "`r params$date`"
output:
  pdf_document:
    latex_engine: xelatex
geometry: "a4paper, margin=2.5cm"
fontsize: 12pt
params:
  title: "Cashflow analysis"
  org: "Organisation name"
  author: "Your name FIA"
  date: "September 2019"
---

```{r setup, include=FALSE}
# Knitr setup
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir= normalizePath('..'))
```

```{r include=FALSE}
# Load project
library("ProjectTemplate")
load.project()

# Run code in analysis.R script
source("src/analysis.R")
```

## Summary

Using a discount rate of `r format(100*disc)`% p.a. the present value of the projected cashflows is &pound;`r format(round(pv_central,-3), big.mark=",")`.

95% of model outcomes have a present value in the range &pound;`r format(round(pv_lower,-3), big.mark=",")` to &pound;`r format(round(pv_upper,-3), big.mark=",")`.

## Cashflow analysis

Lorem ipsum dolor sit amet, ad mea sumo vocibus graecis, at mea soleat doctus, usu elit dicta ne. Aliquid
~~~

Cashflow analysis

Organisation name

September 2019

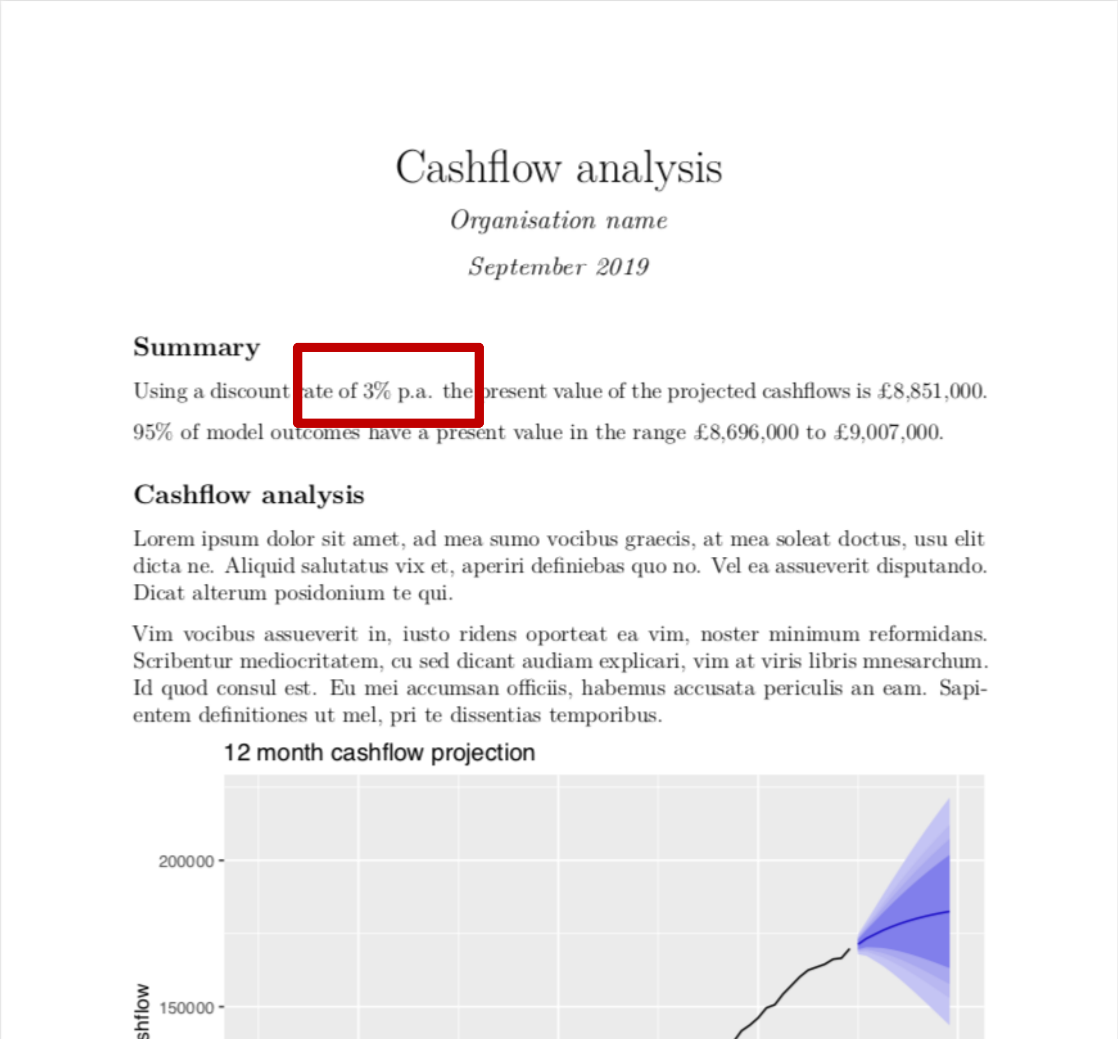
a. the present value of th s is £8,851,000.

a present value in the range £8,696,000 to £9,007,000.

ad mea sumo vocibus graecis, at mea soleat doctus, usu elit x et, aperiri definiebas quo no. Vel ea assueverit disputando. qui.

isto ridens oporteat ea vim, noster minimum reformidans. sed dicant audiam explicari, vim at viris libris mnesarchum. accumsan officiis, habemus accusata periculis an eam. Sapite dissentias temporibus.

v projection



Cashflow analysis

Organisation name

September 2019

Summary

Using a discount rate of 3% p.a. the present value of the projected cashflows is £8,851,000. 95% of model outcomes have a present value in the range £8,696,000 to £9,007,000.

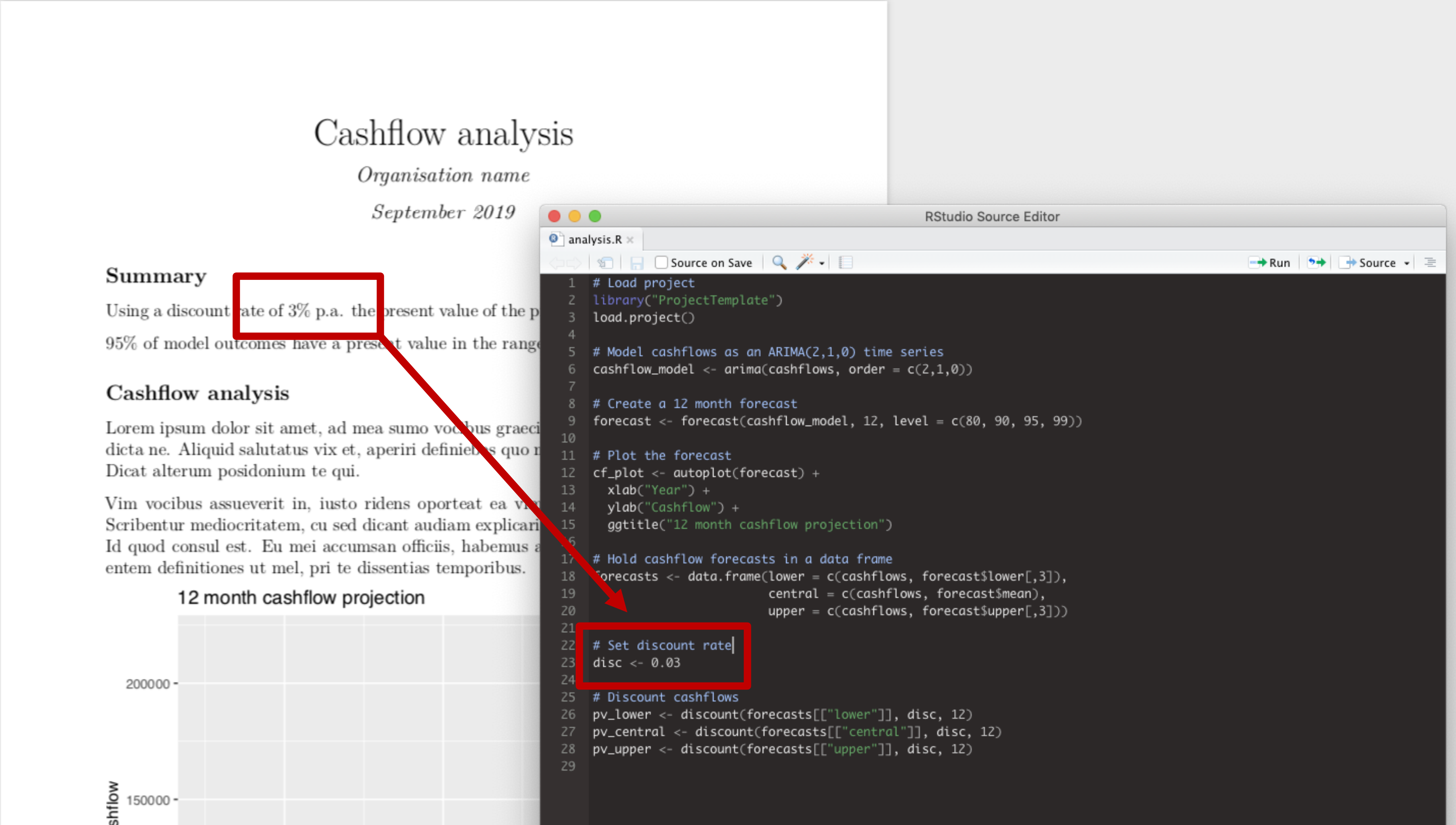
Cashflow analysis

Lorem ipsum dolor sit amet, ad mea sumo vocibus graecis, at mea soleat doctus, usu elit dicta ne. Aliquid salutatus vix et, aperiri definiebas quo no. Vel ea assueverit disputando. Dicat alterum posidonium te qui.

Vim vocibus assueverit in, iusto ridens oporteat ea vim, noster minimum reformidans. Scribentur mediocritatem, cu sed dicant audiam explicari, vim at viris libris mnesarchum. Id quod consul est. Eu mei accumsan officiis, habemus accusata periculis an eam. Sapientem definitiones ut mel, pri te dissentias temporibus.

12 month cashflow projection

> “Update the report using 2.75%”



Cashflow analysis

Organisation name

September 2019

Summary

Using a discount rate of 3% p.a. the present value of the p

95% of model outcomes have a present value in the range

Cashflow analysis

Lorem ipsum dolor sit amet, ad mea sumo vocibus graeci dicta ne. Aliquid salutatus vix et, aperiri definiebas quo r Dicat alterum posidonium te qui.

Vim vocibus assueverit in, iusto ridens oporteat ea v Scribentur mediocritatem, cu sed dicant audiam explicari Id quod consul est. Eu mei accumsan officiis, habemus a entem definitiones ut mel, pri te dissentias temporibus.

```r
# Load project
library("ProjectTemplate")
load.project()

# Model cashflows as an ARIMA(2,1,0) time series
cashflow_model <- arima(cashflows, order = c(2,1,0))

# Create a 12 month forecast
forecast <- forecast(cashflow_model, 12, level = c(80, 90, 95, 99))

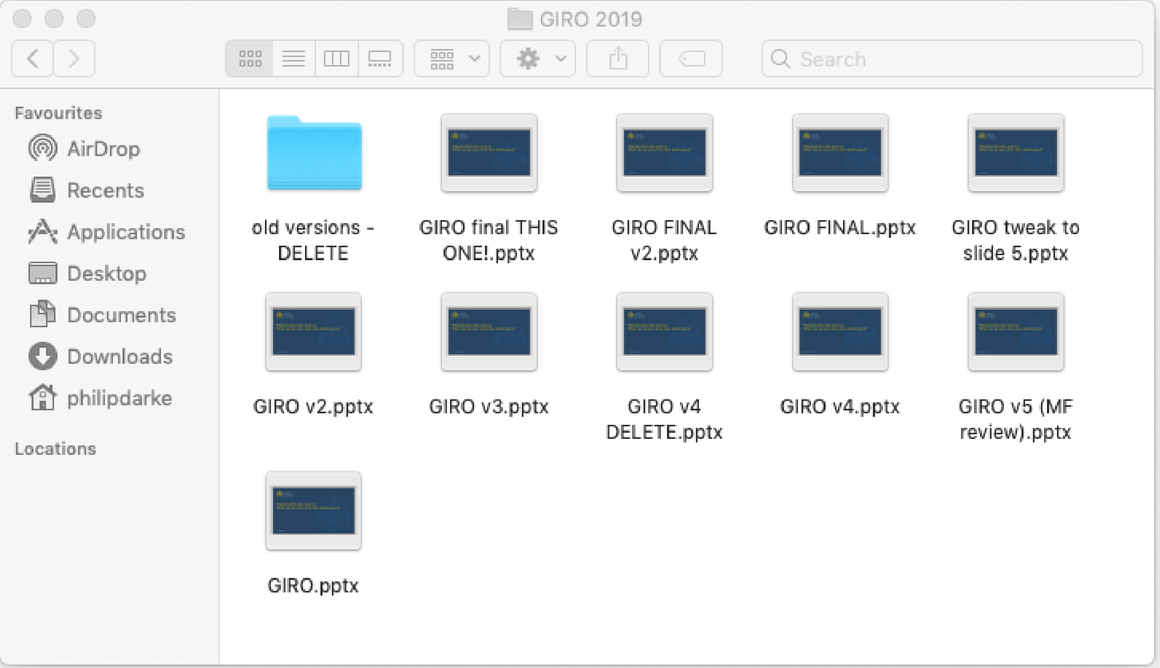
# Plot the forecast
cf_plot <- autoplot(forecast) +
  xlab("Year") +
  ylab("Cashflow") +
  ggtitle("12 month cashflow projection")

# Hold cashflow forecasts in a data frame
forecasts <- data.frame(lower = c(cashflows, forecast$lower[,3]),
                        central = c(cashflows, forecast$mean),
                        upper = c(cashflows, forecast$upper[,3]))

# Set discount rate
disc <- 0.03

# Discount cashflows
pv_lower <- discount(forecasts[["lower"]], disc, 12)
pv_central <- discount(forecasts[["central"]], disc, 12)
pv_upper <- discount(forecasts[["upper"]], disc, 12)
```

Collaboration and keeping an audit trail



GIRO 2019

old versions - DELETE

GIRO final THIS ONE!.pptx

GIRO FINAL v2.pptx

GIRO FINAL.pptx

GIRO tweak to slide 5.pptx

GIRO v2.pptx

GIRO v3.pptx

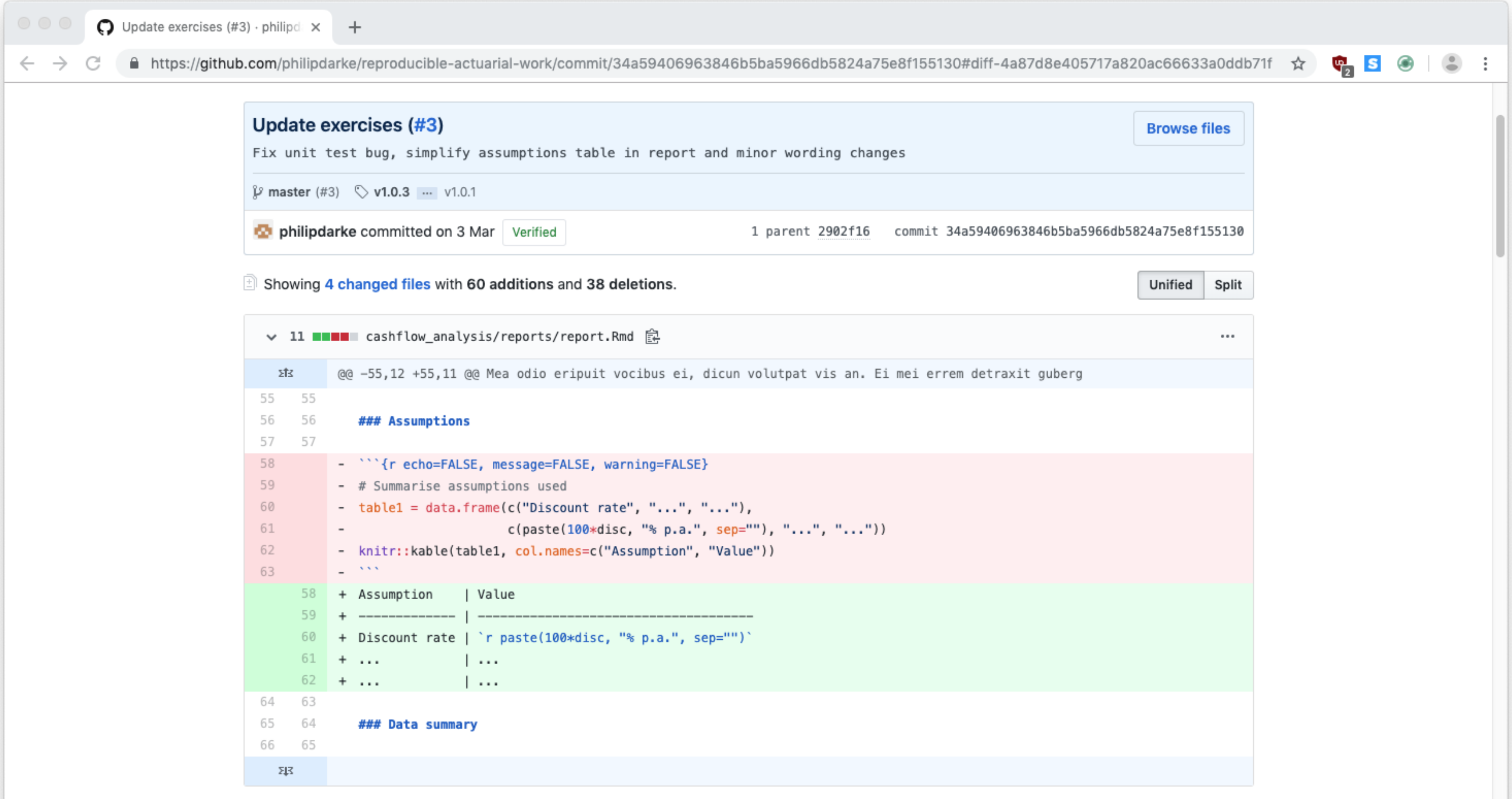
GIRO v4 DELETE.pptx

GIRO v4.pptx

GIRO v5 (MF review).pptx

GIRO.pptx

Collaboration and keeping an audit trail



Update exercises (#3) · philipd
https://github.com/philipdarke/reproducible-actuarial-work/commit/34a59406963846b5ba5966db5824a75e8f155130#diff-4a87d8e405717a820ac66633a0ddb71f
Browse files
Unified
Split

Update exercises (#3)

```
Fix unit test bug, simplify assumptions table in report and minor wording changes
```

master (#3) v1.0.3 ... v1.0.1

philipdarke committed on 3 Mar Verified 1 parent 2902f16 commit 34a59406963846b5ba5966db5824a75e8f155130

Showing 4 changed files with 60 additions and 38 deletions.

11 cashflow_analysis/reports/report.Rmd

```
@@ -55,12 +55,11 @@ Mea odio eripuit vocibus ei, dicun volutpat vis an. Ei mei errem detraxit guberg

  ### Assumptions

- ```{r echo=FALSE, message=FALSE, warning=FALSE}
- # Summarise assumptions used
- table1 = data.frame(c("Discount rate", "...", "..."),
-                     c(paste(100*disc, "% p.a.", sep=""), "...", "..."))
- knitr::kable(table1, col.names=c("Assumption", "Value"))
- ```
+ Assumption    | Value
+ ------------- | -------------------------------------
+ Discount rate | `r paste(100*disc, "% p.a.", sep="")`
+ ...           | ...
+ ...           | ...

  ### Data summary

```

Challenges

- Relies on open source software
- Timing consuming to set up
- Training requirements

Institute and Faculty of Actuaries

Build a simple reproducible pipeline at

philipdarke.com/reproducible-actuarial-work

Applying these techniques in your work

- Take an existing process
- Develop a minimal viable solution (see the exercises)
- Pilot it and let others contribute
- Share what you learn

Institute
and Faculty
of Actuaries

Questions

Comments

The views expressed in this presentation are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the authors.

Institute
and Faculty
of Actuaries

Useful tools for building a reproducible workflow

RStudio is a free and widely used development environment for R that integrates with the tools below.

ProjectTemplate automates the menial parts of statistical analysis and provides a standard way of working in R.

R Markdown is a notebook interface that allows code to sit alongside narrative text and can be used for reporting as part of a reproducible framework with ggplot2 for creating charts and visualisations.

Git is a version control system for managing code and audit trails – it can be used privately in an organisation or with a web-based service such as GitHub.

testthat is a formal automated testing (“unit testing”) package for R.

TravisCI integrates with GitHub to automatically run your tests when code is updated.

roxygen2 automates the production of documentation for your code in R.

Docker packages dependencies inside a container which can run consistently on any infrastructure (also see checkpoint/packrat or consider creating a R package).

References and resources

- RAP companion https://ukgovdatascience.github.io/rap-website/
- RAP Udemy video course https://www.udemy.com/course/reproducible-analytical-pipelines/
- Blog post on the use of R at the BBC https://medium.com/bbc-visual-and-data-journalism/how-the-bbc-visual-and-data-journalism-team-works-with-graphics-in-r-ed0b35693535
- Accompanying exercises https://philipdarke.com/reproducible-actuarial-work/
- Icons made by Smashicons and Dimitry Miroliubov from www.flaticon.com

Get in touch

Dr Matthew Forshaw is a Lecturer in Data Science at Newcastle University, and Data Skills Policy Leader at The Alan Turing Institute working on the Data Skills Taskforce. He is the Programme Director of Newcastle's Industrial MSc in Data Science.
mattforshaw.com

Philip Darke is an actuary with over 10 years' consulting experience at Mercer and a PhD researcher in data science at the EPSRC Centre for Doctoral Training in Cloud Computing for Big Data at Newcastle University.
philipdarke.com

Institute
and Faculty
of Actuaries