

The lazy and easily distracted report writer

Mike K Smith (Pfizer R&D UK Ltd.)
@MikeKSmith

RStudio::conf(2019)

TL;DR - Disclaimer

- I used rmarkdown notebooks / documents to write up an exploratory analysis to share with a drug development team.
 - Statistician
 - Clinical Pharmacologist
 - Clinician
 - My manager
- The analysis presented here is **NOT** that analysis (for confidentiality) but it has similar attributes.

Your (my) brain is
lazy, shallow, and
easily distracted.

<https://www.slideshare.net/CJAtherton/chris-atherton-at-presentation-camp-london>

Cutlery drawers & what they say about YOU



HT: @HadleyWickham, @jimhester_, @dataandme

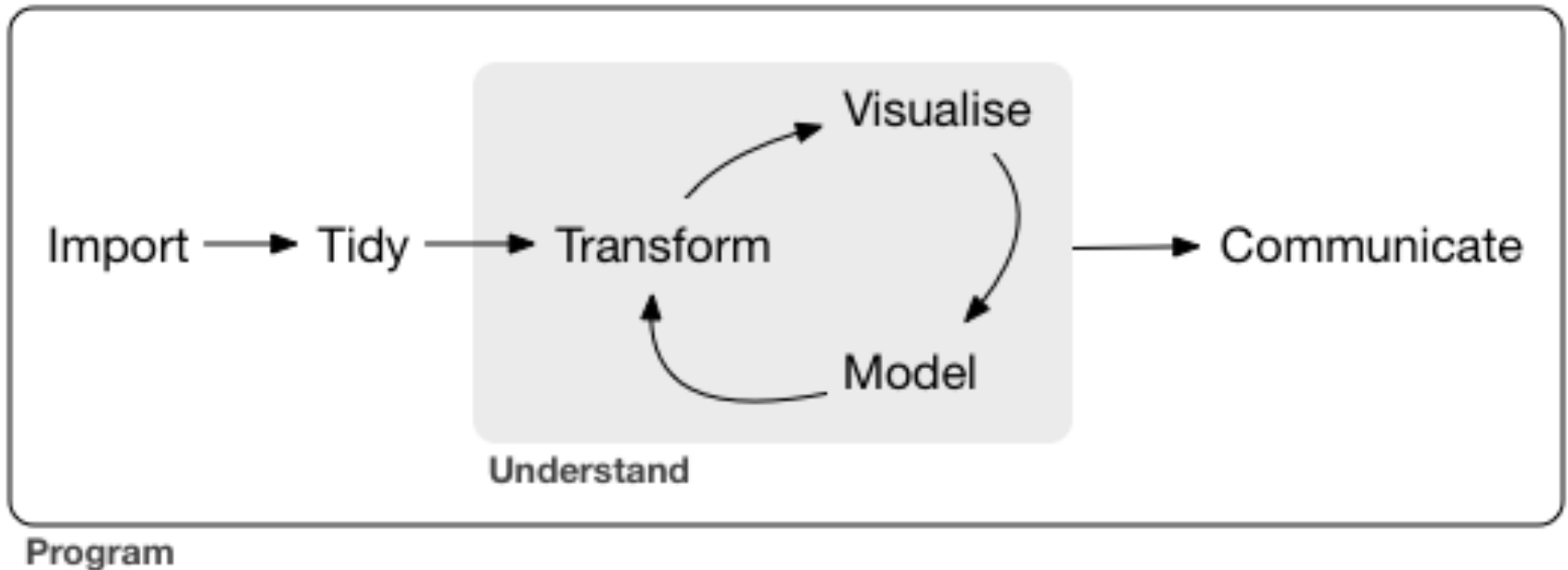
Mine... (sorry / not sorry)





HOME ORGANIZATION TIP:
JUST GIVE UP.

Data analysis - THEORY



<https://r4ds.had.co.nz/>

Data Analysis - In practice....

(**DISCLAIMER**: I'm **sure** the experiences recounted here are **unique to me alone**.
See cutlery drawer pictures for proof...)

Go to email with link to data
source...

read and respond to 3 other emails...

Download and read data into R...

*stop and answer colleague's question(s)
about the tidyverse...*

Wrangle data and plot it...

LUNCH

*go to an (unrelated) meeting /
teleconference call.*

Make better plots.

*follow an interesting link that
Mara Averick (@dataandme)
just posted on Twitter*

Fit preliminary model to data.

<Next day>

Team find problem with data,
share new version of data.

Change input data and redo
analysis.

Check new version against
previous version.

Discuss findings with my boss.

file expenses.

Circulate report.

DONE!!!

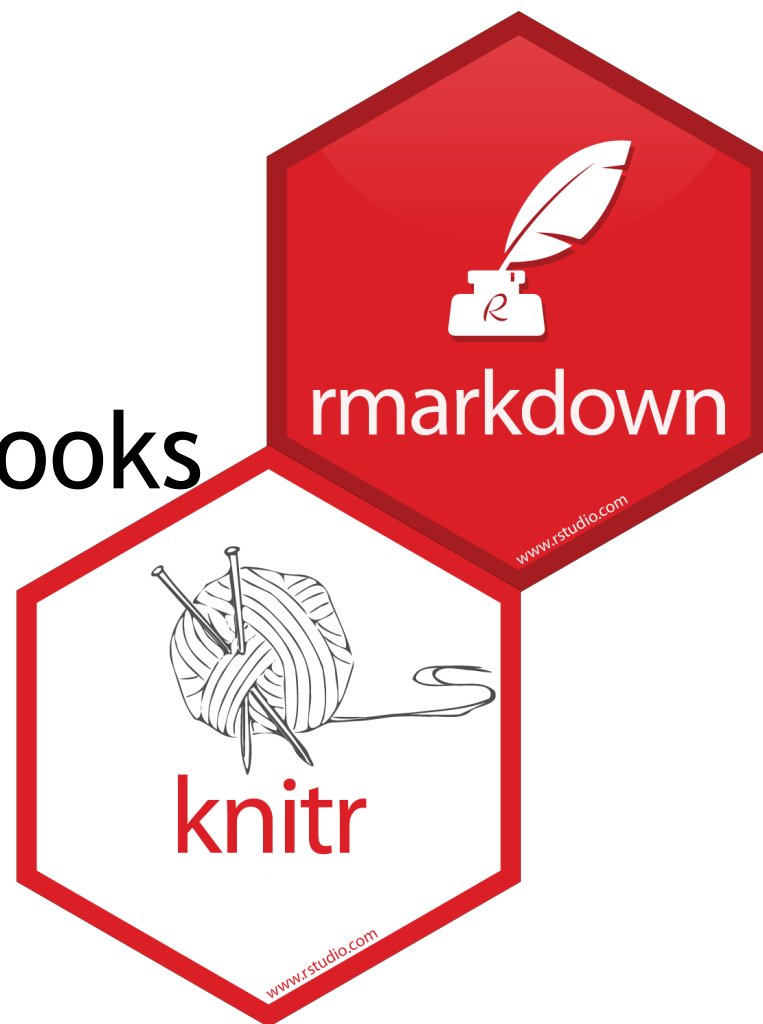
< 3 months pass >

Review comments come back...

Wait... Erm... **WHAT** was I thinking?

To the rescue...

rmarkdown & notebooks



Who is your audience?

- Present (distracted) *me*
- Future (3 months later) *me*
- Quantitative colleagues / reviewers
- Decision makers (may not be quantitative)

Notebooks / markdown vs scripts *(for analysis)*



Mike K Smith @MikeKSmith · Sep 12

My opinion: If you write more comments (explanation) than code, use markdown. If you write more code than comments, then write more comments and use markdown. @StatGarrett #earlconf



1



19



101



Show this thread

BUT, see also: <https://yihui.name/en/2018/09/notebook-war/>

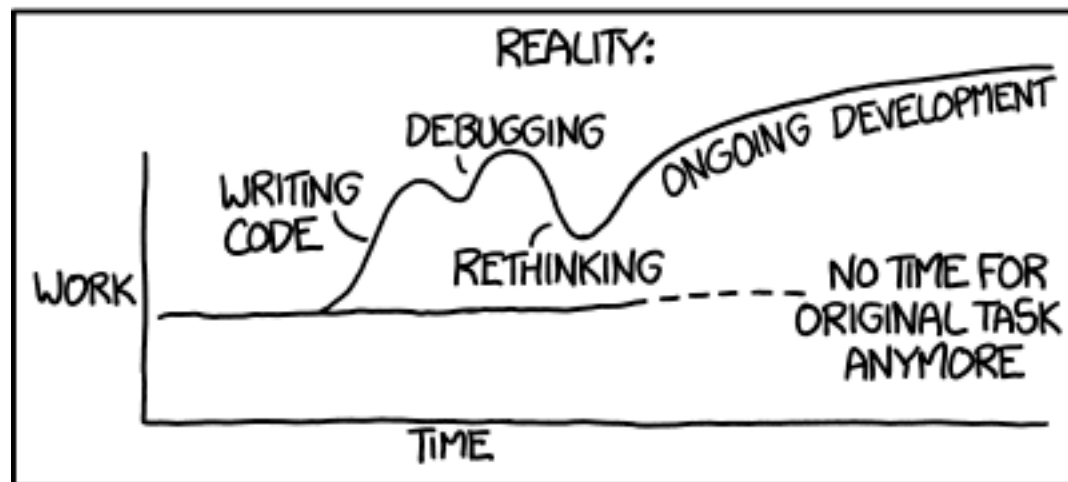
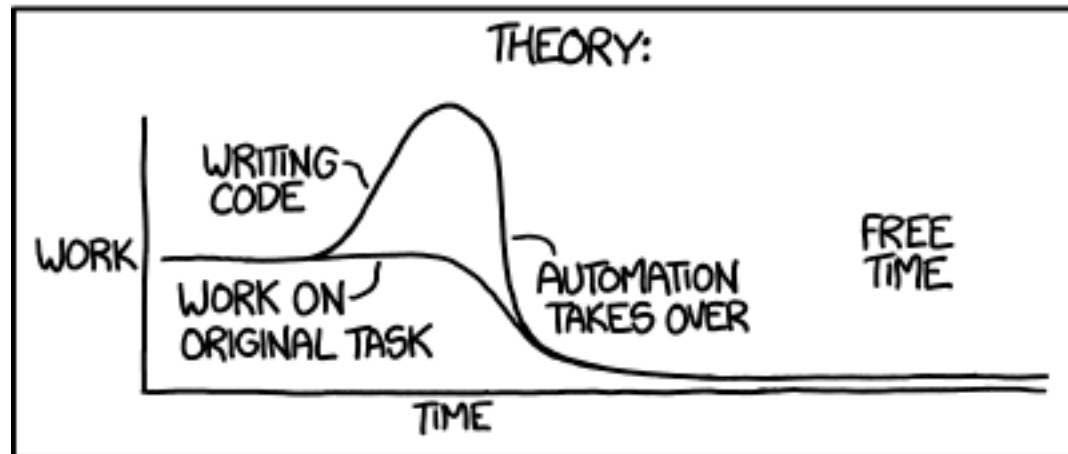
Also...

I *knew* my manager / other reviewers
would ask for reports
on the *THREE* different endpoints.

Rule of three

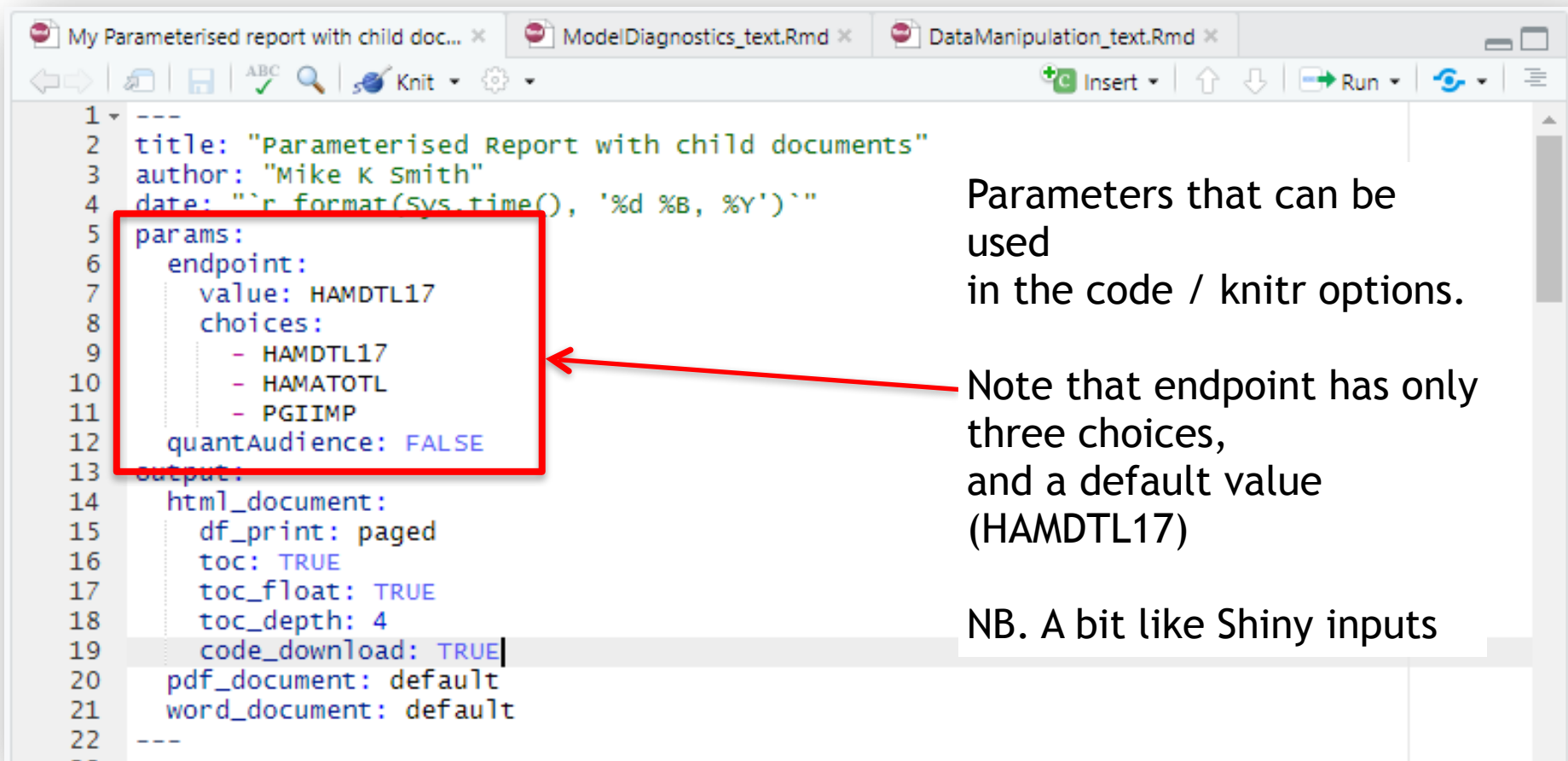
- Copy & paste code ≥ 3 times?
 - Write and use a function
- Perform analysis across ≥ 3 endpoints?
 - Multiple markdown reports?
 - ***NOPE. Parameterised*** reports.

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



<https://xkcd.com/1319/>

YAML header parameters



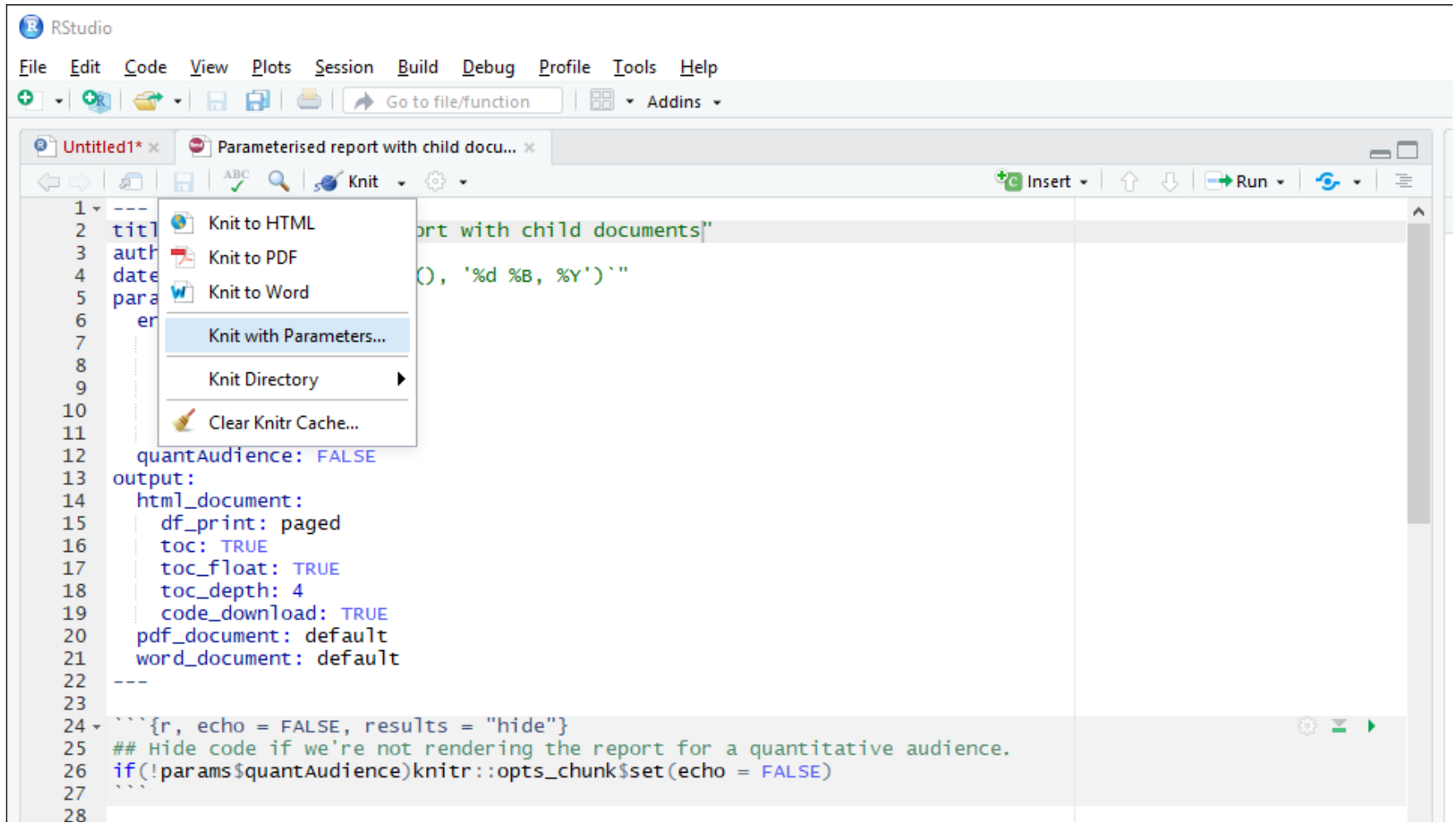
```
1 ---
2 title: "Parameterised Report with child documents"
3 author: "Mike K Smith"
4 date: "`r format(Sys.time(), '%d %B, %Y')`"
5 params:
6   endpoint:
7     value: HAMDTL17
8     choices:
9       - HAMDTL17
10      - HAMATOTL
11      - PGIIMP
12   quantAudience: FALSE
13 output:
14   html_document:
15     df_print: paged
16     toc: TRUE
17     toc_float: TRUE
18     toc_depth: 4
19     code_download: TRUE
20   pdf_document: default
21   word_document: default
22 ---
```

Parameters that can be used in the code / knitr options.

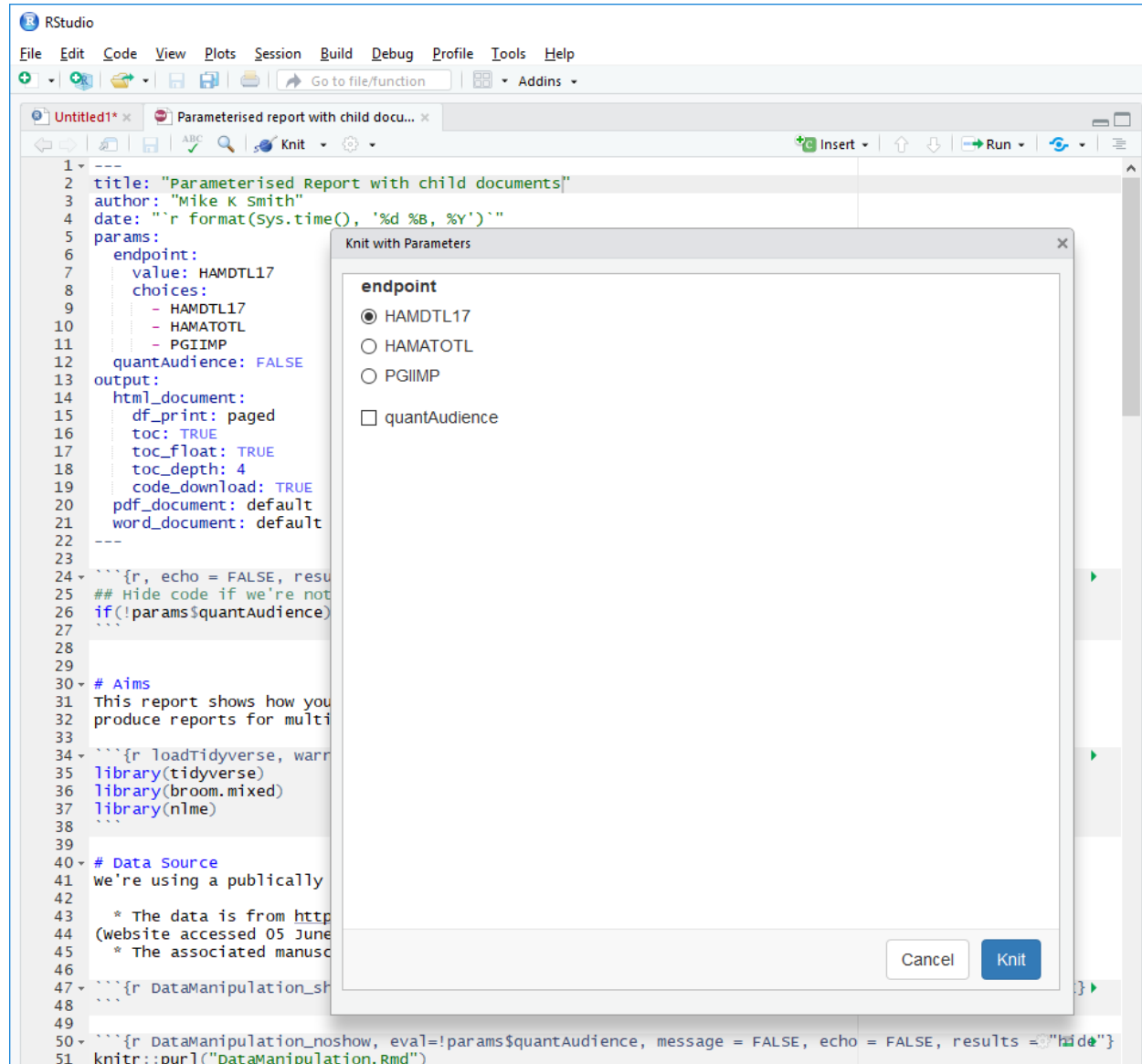
Note that endpoint has only three choices, and a default value (HAMDTL17)

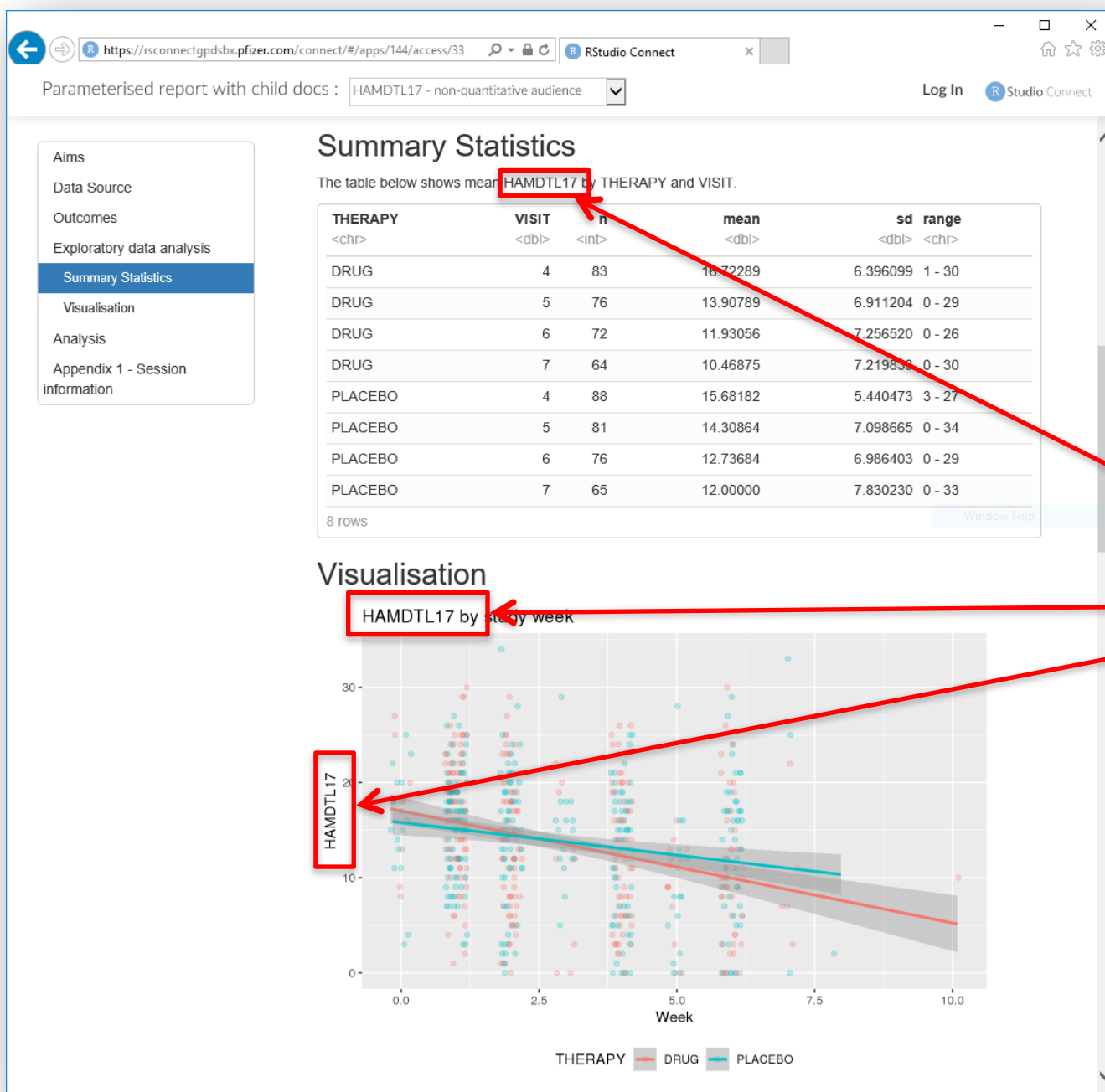
NB. A bit like Shiny inputs

Render with parameters



Render with parameters





params :

endpoint: **HAMDTL17**

quantAudience: **FALSE**

Using
params\$endpoint
in markdown text,
plot code

```
My Parameterised report with child doc... x ModelDiagnostics_text.Rmd x DataManipulation_text.Rmd x
< > ABC 🔍 Knit ⚙️ Insert ↑ ↓ Run ↻
1 ---
2 title: "Parameterised Report with child documents"
3 author: "Mike K Smith"
4 date: "`r format(Sys.time(), '%d %B, %Y')`"
5 params:
6   endpoint:
7     value: HAMDTL17
8     choices:
9       - HAMDTL17
10      - HAMATOTL
11      - PGIIMP
12   quantAudience: FALSE
13 output:
14   html_document:
15     df_print: paged
16     toc: TRUE
17     toc_float: TRUE
18     toc_depth: 4
19     code_download: TRUE
20   pdf_document: default
21   word_document: default
22 ---
23
24 ```{r, echo = FALSE, results = "hide"}
25 ## Hide code if we're not rendering the report for a quantitative audience.
26 if(!params$quantAudience)knitr::opts_chunk$set(echo = FALSE)
27 ```
28
```

Parameters that can be used
in the code / knitr options

Knitr options:
Show code in output **ONLY IF** quantitative audience,

Rename endpoint variable(s) to “outcome” (simplifies later code)

```
{r DataManipulation, results="hide", message=FALSE, warnings=FALSE}
data <- haven::read_sas("chapter15_example.sas7bdat")

data <- data %>%
  rename_all(funs(
    str_replace(string = ., pattern=params$endpoint, replacement="outcome")
  )) %>%
  bind_cols(data,.) %>%
  drop_na()
```

Run this code **ONLY IF** `params$quantAudience = TRUE`

```
{r Show_data, eval = params$quantAudience}
data %>%
  head(10)
```

Run **ONLY IF** `params$quantAudience = TRUE`, to pull in text *from child document*

```
{r DataManipulationChildDoc, eval=params$quantAudience,
  child="DataManipulation_text.Rmd"}
```


https://rsconnectgdsbx.pfizer.com/connect/#/apps/144/access/33 RStudio Connect

Content / Parameterised report with child docs : HAMDTL17 - quantitative audience

Data Source

We're using a publicly available dataset on depression.

- The data is from <https://missingdata.lshtm.ac.uk/category/dia-working-group/example-data-sets/> (Website accessed 05 June 2018).
- The associated manuscript is <https://www.ncbi.nlm.nih.gov/pubmed/15232330>.

```
data <- haven::read_sas("chapter15_example.sas7bdat")

data <- data %>%
  rename_all(funs(
    str_replace(string = ., pattern=params$endpoint, replacement="outcome")
  )) %>%
  bind_cols(data,.) %>%
  drop_na()
```

```
data %>%
  head(10)
```

PATIENT	HAMATOTL	PGIIMP	RELDAYS	VISIT	THERAPY	GEND...	POOLINV	basval
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>
1503	21	2	7	4	DRUG	F	006	32
1503	19	2	14	5	DRUG	F	006	32
1503	21	3	28	6	DRUG	F	006	32
1503	17	4	42	7	DRUG	F	006	
1507	18	3	7	4	PLACEBO	F	006	14
1507	18	2	15	5	PLACEBO	F	006	14
1507	14	3	29	6	PLACEBO	F	006	14
1507	8	2	42	7	PLACEBO	F	006	14
1509	18	3	7	4	DRUG	F	006	21
1509	17	3	14	5	DRUG	F	006	21

1-10 of 10 rows | 1-9 of 22 columns

Data Manipulations

The data shows post-baseline measurements for subjects on duloxetine and placebo (+paroxetine). Although two doses of duloxetine were given in the study the data has been "anonymised" by randomly sampling from the two different doses.

Conditional execution of Data Manipulation.

params:

endpoint: HAMDTL17

quantAudience: **TRUE**

Code shown

Chunk run

Data shown

Child doc text

https://rsconnectgdsbx.pfizer.com/connect/#/apps/144/access/33 RStudio Connect

Content / Parameterised report with child docs

HAMDTL17 - quantitative audience

Data Source

We're using a publicly available dataset on depression.

- The data is from <https://missingdata.lshtm.ac.uk/category/dia-working-group/example-data-sets/> (Website accessed 05 June 2018).
- The associated manuscript is <https://www.ncbi.nlm.nih.gov/pubmed/15232330>.

```
data <- haven::read_sas("chapter15_example.sas7bdat")

data <- data %>%
  rename_all(funs(
    str_replace(string = ., pattern=params$endpoint, replacement="outcome")
  )) %>%
  bind_cols(data, .) %>%
  drop_na()

data %>%
  head(10)
```

PATIENT	HAMATOTL	PGIIMP	RELDAYS	VISIT	THERAPY	GEND...	POOLINV	basval
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>
1503	21	2	7	4	DRUG	F	006	32
1503	19	2	14	5	DRUG	F	006	32
1503	21	3	28	6	DRUG	F	006	32
1503	17	4	42	7	DRUG	F	006	32
1507	18	3	7	4	PLACEBO	F	006	14
1507	18	2	15	5	PLACEBO	F	006	14
1507	14	3	29	6	PLACEBO	F	006	14
1507	8	2	42	7	PLACEBO	F	006	14
1509	18	3	7	4	DRUG	F	006	21
1509	17	3	14	5	DRUG	F	006	21

1-10 of 10 rows | 1-9 of 22 columns

Data Manipulations

The data shows post-baseline measurements for subjects on duloxetine and placebo (+paroxetine). Although two doses of duloxetine were given in the study the data has been "anonymised" by randomly sampling from the two different doses.

Conditional execution of Data Manipulation.

INPUT

params :

endpoint: **HAMDTL17**

quantAudience: **TRUE**




















RStudio Connect allows you (or visitor to your page) to specify parameters and render a parameterised report and then to save that report as a named item.

You can then have pre-rendered reports for various audiences ready to go...

More parameterisation

- Question: how to show correct analysis for non-continuous endpoint?
 - Change analysis type in code depending on `params$endpoint`.
- Question: what to do if something goes wrong in the analysis?
 - Check for errors and handle appropriately using `tryCatch(...)`
 - Insert child document text: “**EMERGENCY!** Something has gone wrong... Contact your data scientist!”

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE
EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)

		HOW OFTEN YOU DO THE TASK					
		50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
HOW MUCH TIME YOU SHAVE OFF	1 SECOND	 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
	5 SECONDS	 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
	30 SECONDS	 4 WEEKS	 3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
	1 MINUTE	 8 WEEKS	 6 DAYS	 1 DAY	4 HOURS	1 HOUR	5 MINUTES
	5 MINUTES	9 MONTHS	 4 WEEKS	 6 DAYS	21 HOURS	5 HOURS	25 MINUTES
	30 MINUTES		6 MONTHS	 5 WEEKS	 5 DAYS	 1 DAY	2 HOURS
	1 HOUR		10 MONTHS	2 MONTHS	 10 DAYS	 2 DAYS	5 HOURS
	6 HOURS				2 MONTHS	 2 WEEKS	 1 DAY
	 1 DAY					 8 WEEKS	 5 DAYS

Feel free to ask
me questions,
but remember....

