Data Analysis Pipelines

Data Analysis

Pipelines are **automated** and **reproducible**

Pipelines are helpful when there is a **single question you will have to answer more than once**

Appropriate Checks

- Input file in expected format?
- Necessary variables included in data?
- Observations coded as expected?
- etc...

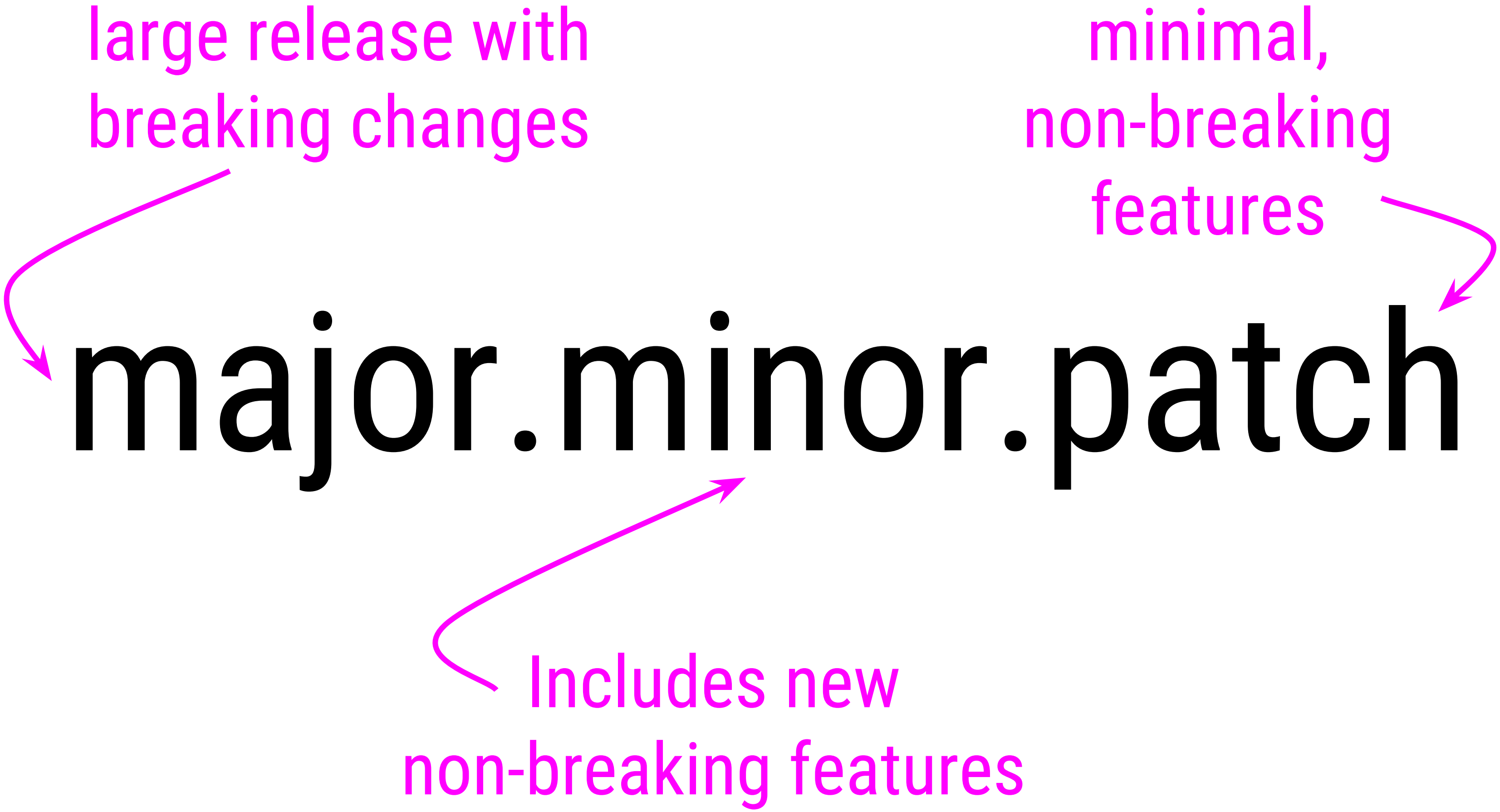
Avoid hard-coding whenever possible

let your code fill the values in for you!

Good pipelines are **scalable** pipelines

Your pipeline should work in the future...when the dataset is likely larger

Pipelines should be versioned

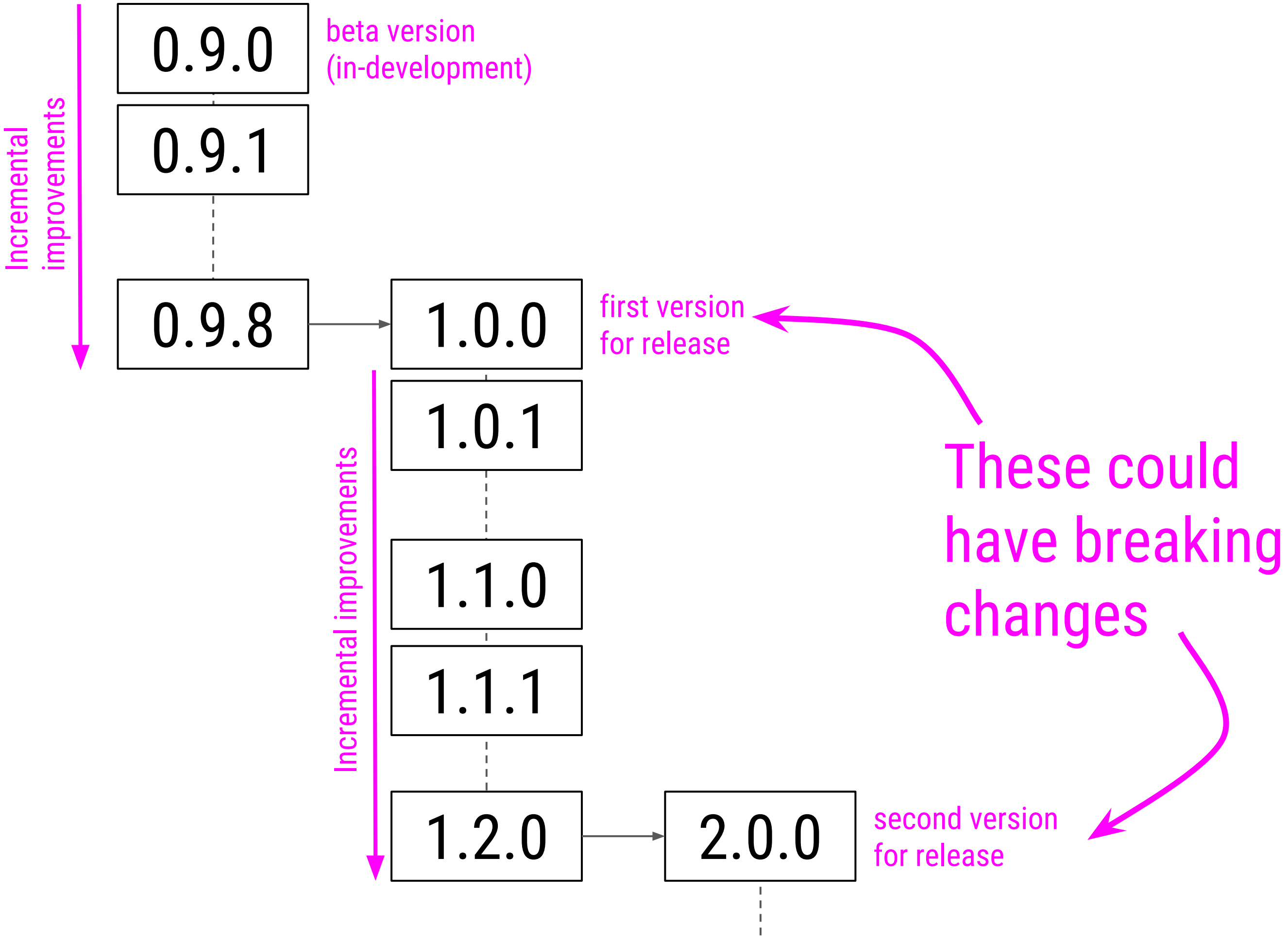


major.minor.patch

large release with breaking changes

Includes new non-breaking features

minimal, non-breaking features



0.9.0

beta version
(in-development)

0.9.1

0.9.8

Incremental
improvements

1.0.0

first version
for release

1.0.1

1.1.0

1.1.1

1.2.0

Incremental improvements

2.0.0

second version
for release

These could have breaking changes

The YAML will specify `params`

```
---
title: My Document
output: html_document
params:
   filename: filename_filedate.csv
---
```

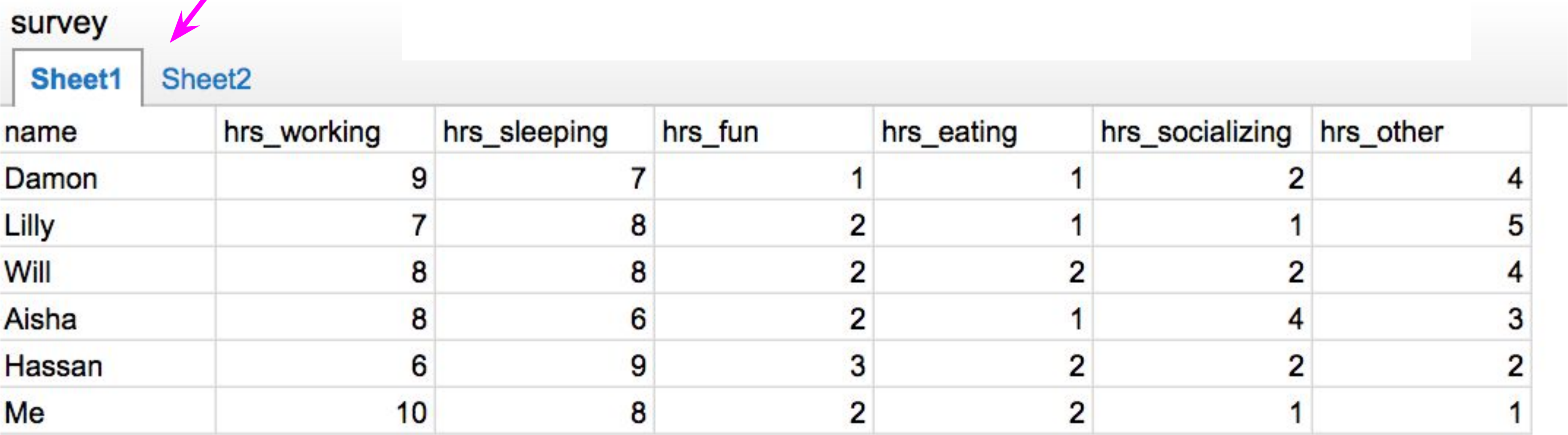
Your YAML will include the parameters you'll use throughout your report

`render()` will knit your parameterized report

```r
rmarkdown::render("MyDocument.Rmd",
params = list(
  filename = "filename_filedate.csv")
)
```

To generate your document,
you can use `render()` and
specify `params` as a list.

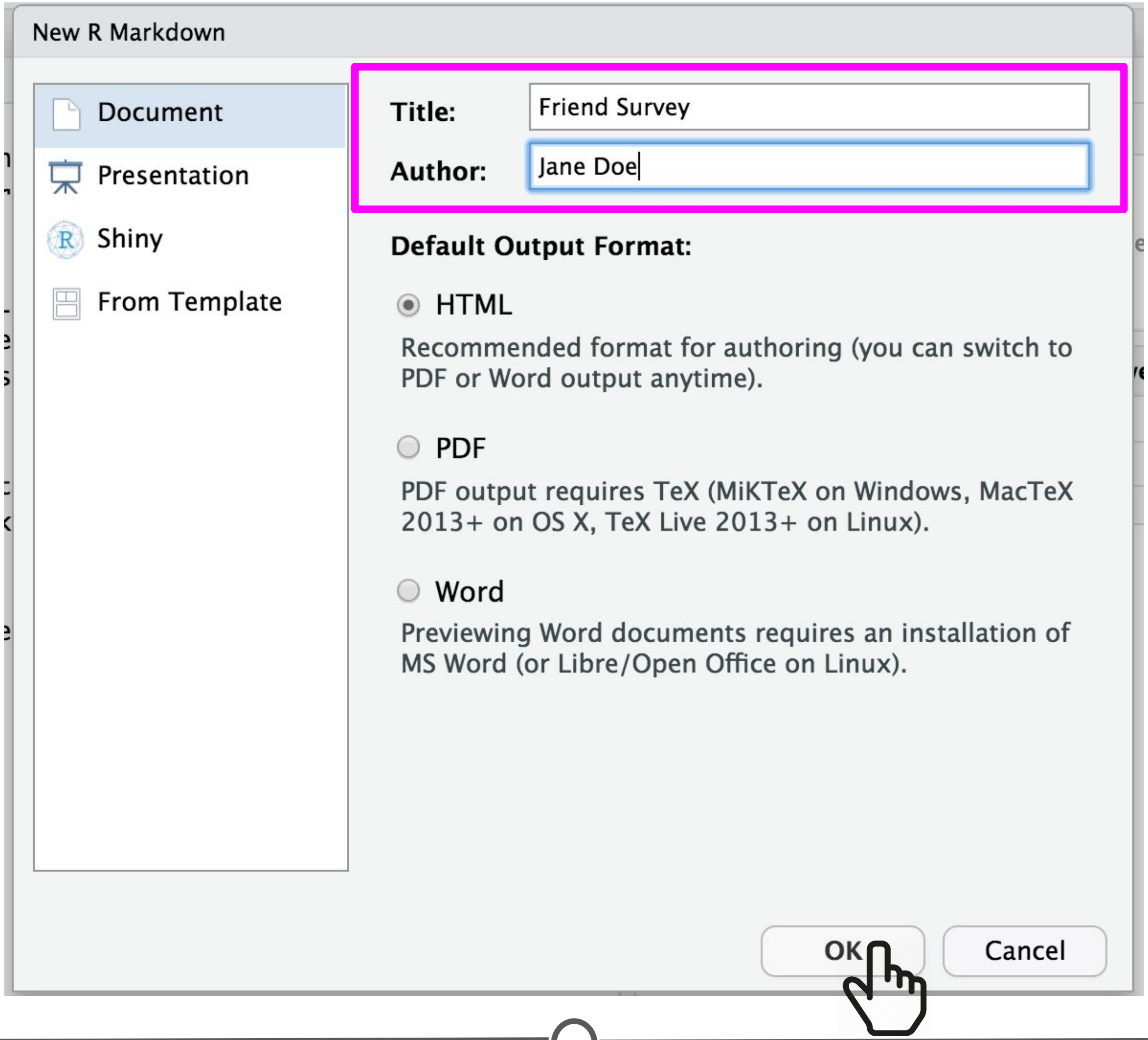
In this example, **Sheet1** is the first week's data, **Sheet2** includes the first two weeks' data



survey

name	hrs_working	hrs_sleeping	hrs_fun	hrs_eating	hrs_socializing	hrs_other
Damon	9	7	1	1	2	4
Lilly	7	8	2	1	1	5
Will	8	8	2	2	2	4
Aisha	8	6	2	1	4	3
Hassan	6	9	3	2	2	2
Me	10	8	2	2	1	1

https://docs.google.com/spreadsheets/d/e/2PACX-1vTuPnBT7Y6emAamDSaXo_RV-X54c0a6nJh4e-bqurRsNW_Elm6ygq90UVfySQbSg0Ue9-U4wSykQqKF/pubhtml



New R Markdown

Title: Friend Survey

Author: Jane Doe

Default Output Format:

● HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

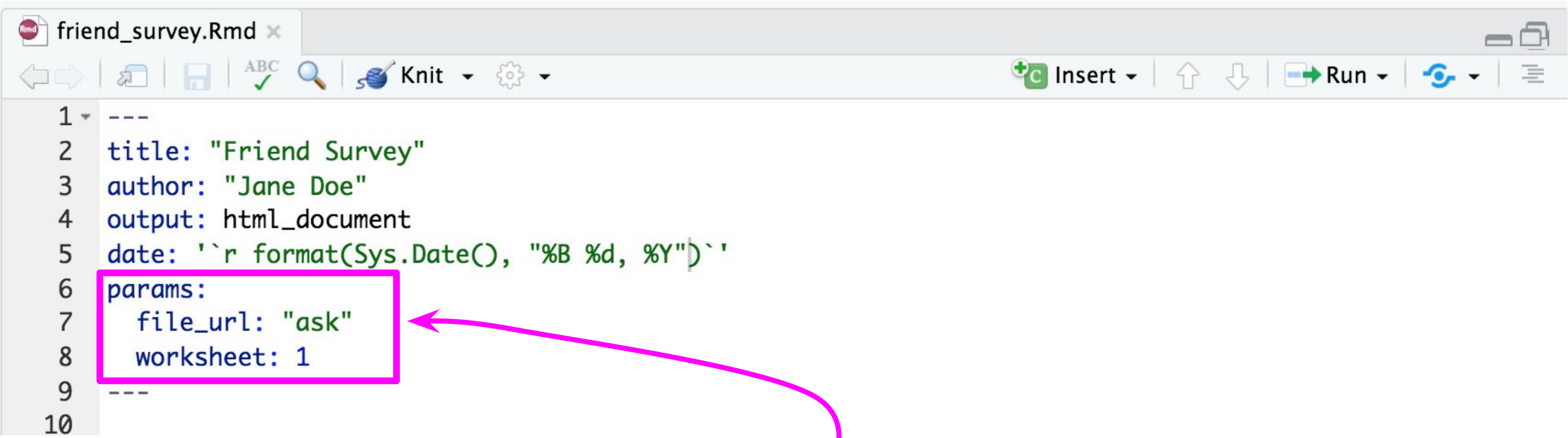
○ PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

We'll start with a typical R Markdown document.

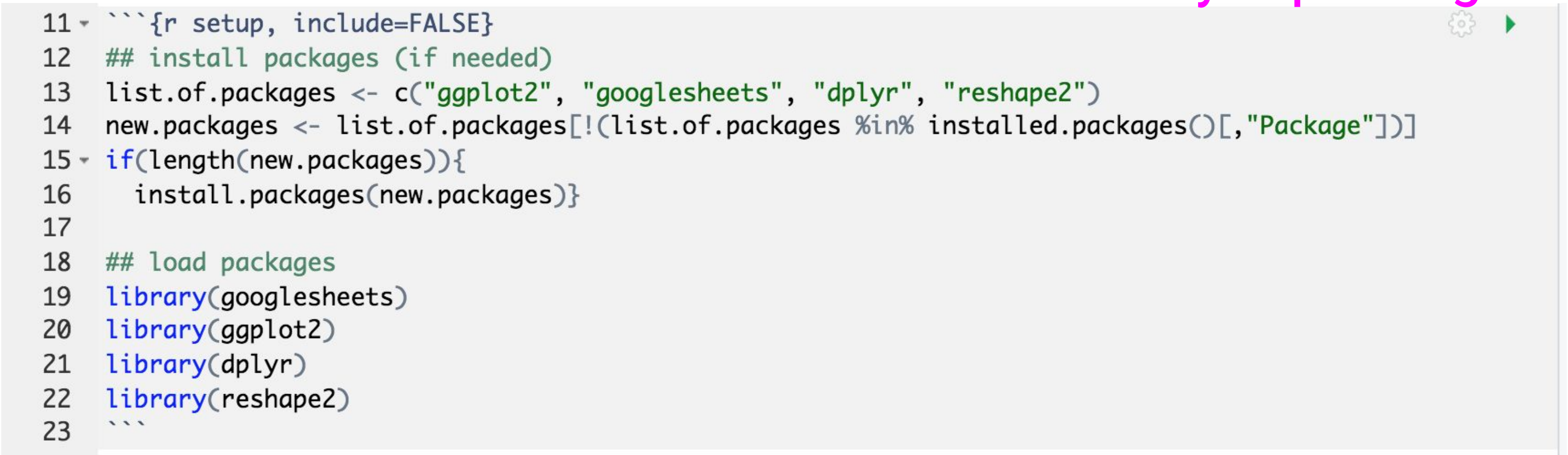


```
---
title: "Friend Survey"
author: "Jane Doe"
output: html_document
date: '`r format(Sys.Date(), "%B %d, %Y")`'
params:
  file_url: "ask"
  worksheet: 1
---

```

`params` are specified within your YAML

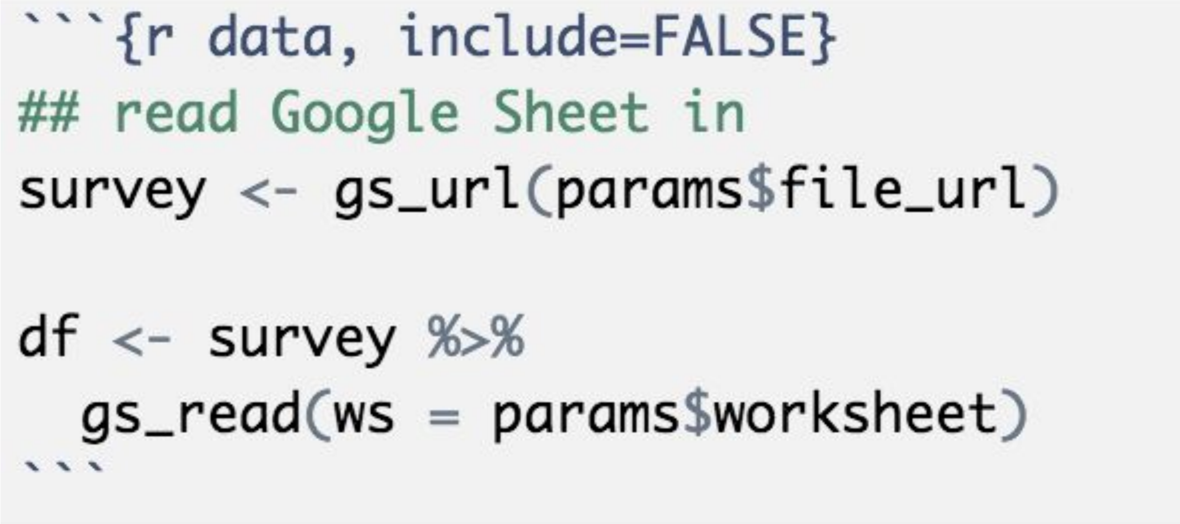
install and load necessary R packages



~~~
```{r setup, include=FALSE}
## install packages (if needed)
list.of.packages <- c("ggplot2", "googlesheets", "dplyr", "reshape2")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)){
  install.packages(new.packages)}

## load packages
library(googlesheets)
library(ggplot2)
library(dplyr)
library(reshape2)
```
~~~

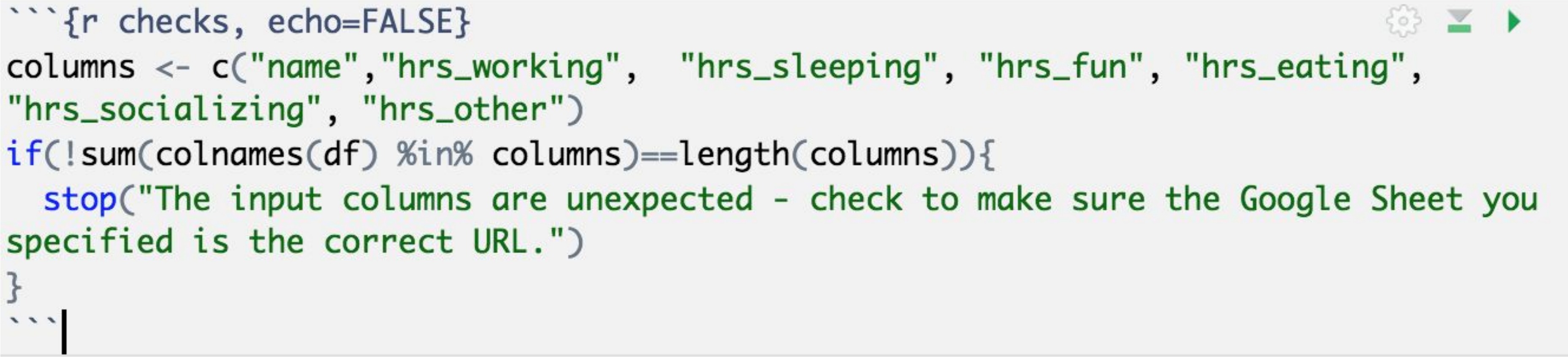
Read in the Google Sheet



~~~
```{r data, include=FALSE}
## read Google Sheet in
survey <- gs_url(params$file_url)

df <- survey %>%
  gs_read(ws = params$worksheet)
```
~~~

Write checks for your data and analysis



~~~
```{r checks, echo=FALSE}
columns <- c("name","hrs_working",  "hrs_sleeping", "hrs_fun", "hrs_eating",
"hrs_socializing", "hrs_other")
if(!sum(colnames(df) %in% columns)==length(columns)){
  stop("The input columns are unexpected - check to make sure the Google Sheet you
specified is the correct URL.")
}
```
~~~

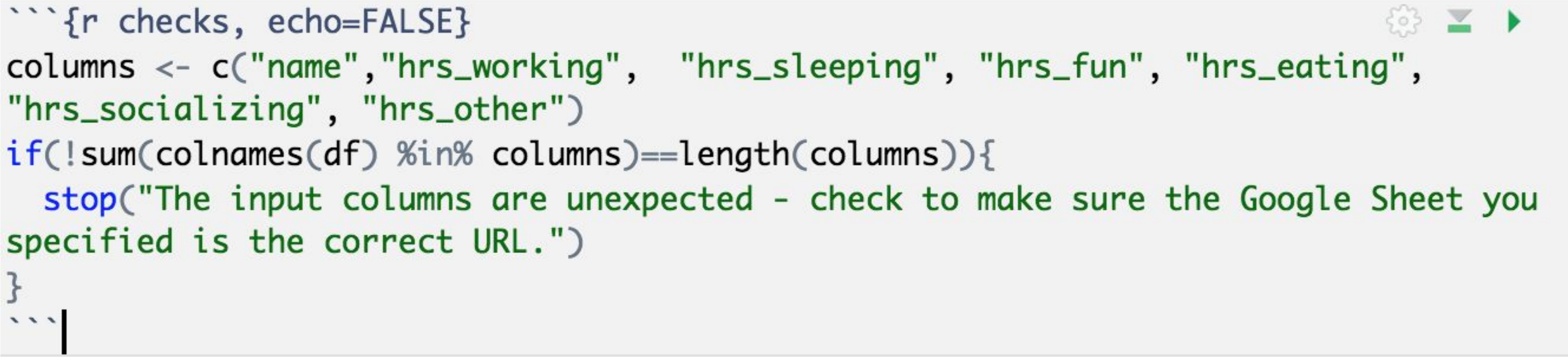
messages, warnings, and stop

`message()` - prints a message, code continues to run

`warning()` - prints a warning, code continues to run

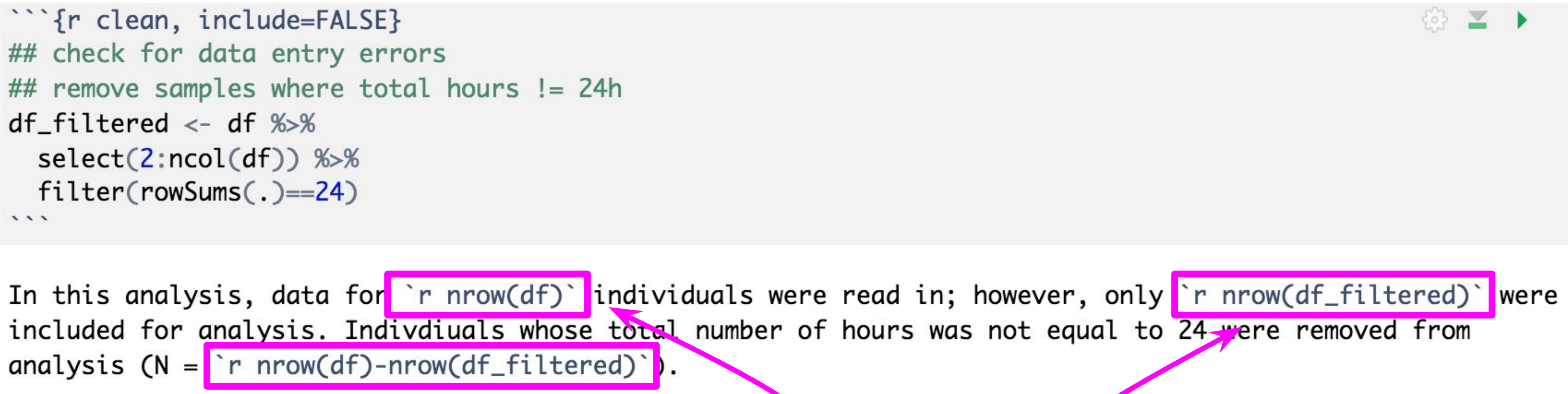
`stop()` - stops code from running, prints error message

Write checks for your data and analysis



~~~
```{r checks, echo=FALSE}
columns <- c("name","hrs_working",  "hrs_sleeping", "hrs_fun", "hrs_eating",
"hrs_socializing", "hrs_other")
if(!sum(colnames(df) %in% columns)==length(columns)){
  stop("The input columns are unexpected - check to make sure the Google Sheet you
specified is the correct URL.")
}
```
~~~

Clean your data!

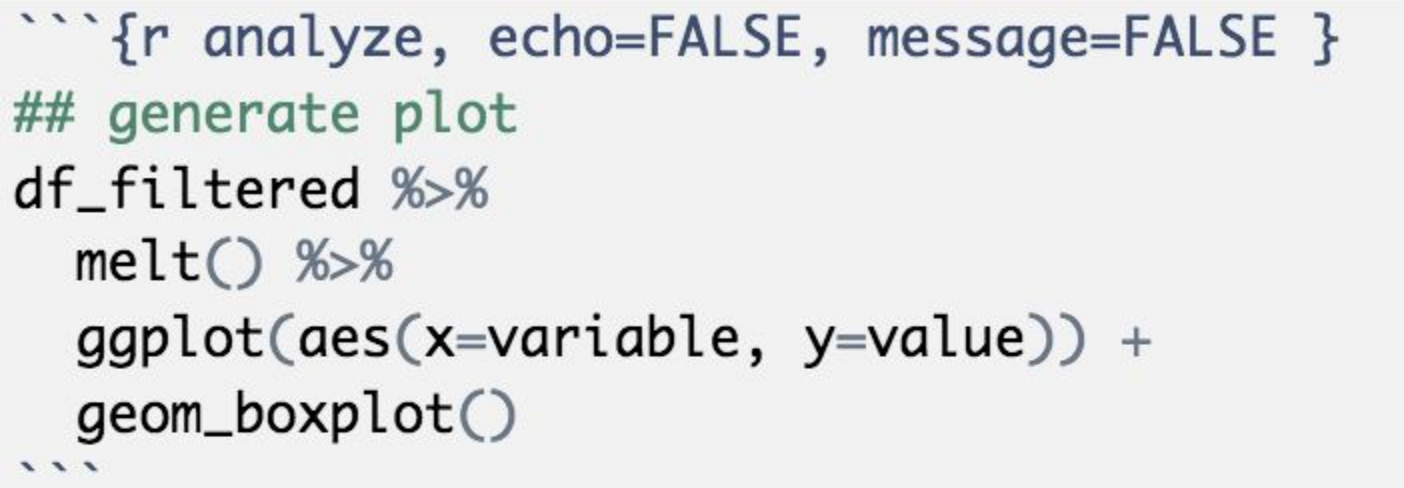


~~~
```{r clean, include=FALSE}
## check for data entry errors
## remove samples where total hours != 24h
df_filtered <- df %>%
  select(2:ncol(df)) %>%
  filter(rowSums(.)==24)
```
~~~

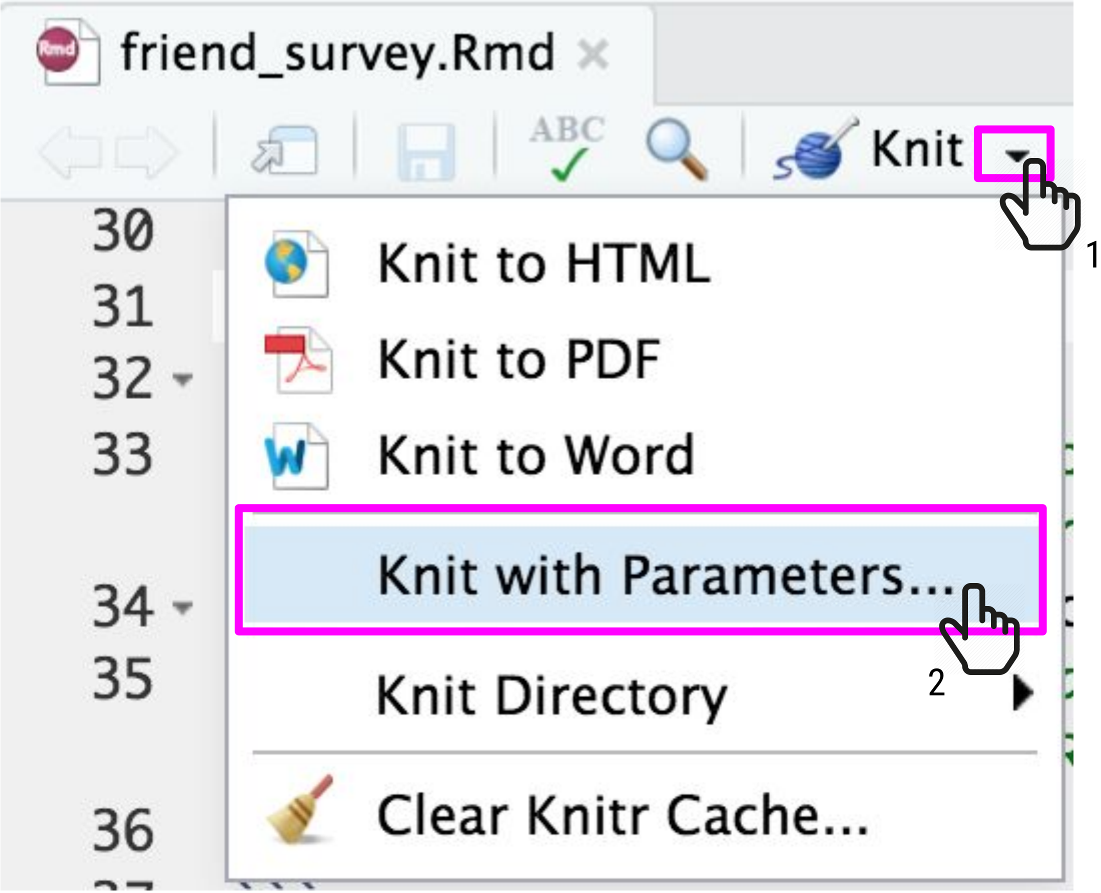
In this analysis, data for `r nrow(df)` individuals were read in; however, only `r nrow(df_filtered)` were included for analysis. Indivdiuals whose total number of hours was not equal to 24 were removed from analysis (N = `r nrow(df)-nrow(df_filtered)`).

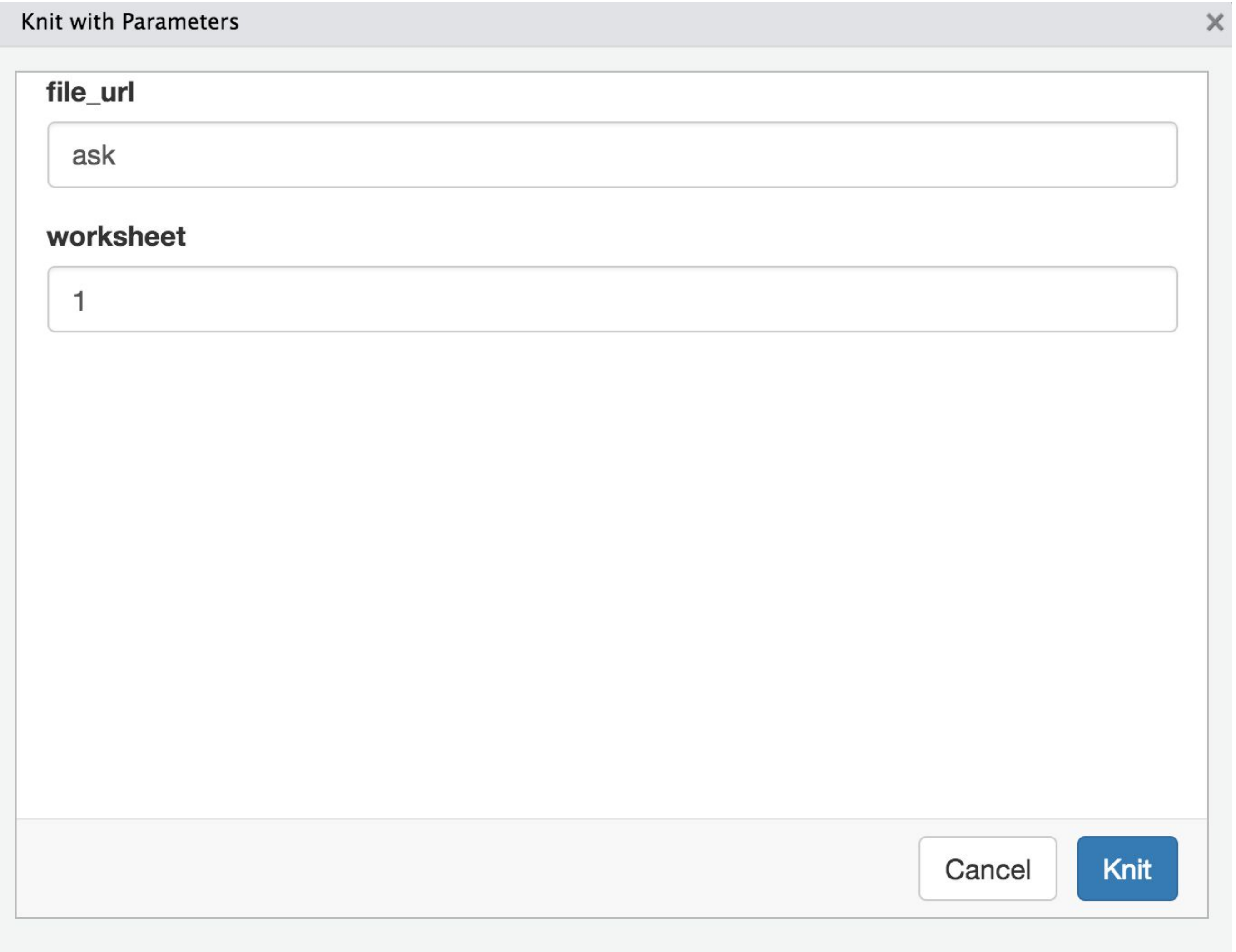
Avoid hard-coding!

Generate plot of interest



~~~~
```{r analyze, echo=FALSE, message=FALSE }
## generate plot
df_filtered %>%
  melt() %>%
  ggplot(aes(x=variable, y=value)) +
  geom_boxplot()
```
~~~~


friend_survey.Rmd
ABC
Knit
1
30
31
32
33
34
35
36
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
2
Knit Directory
Clear Knitr Cache...

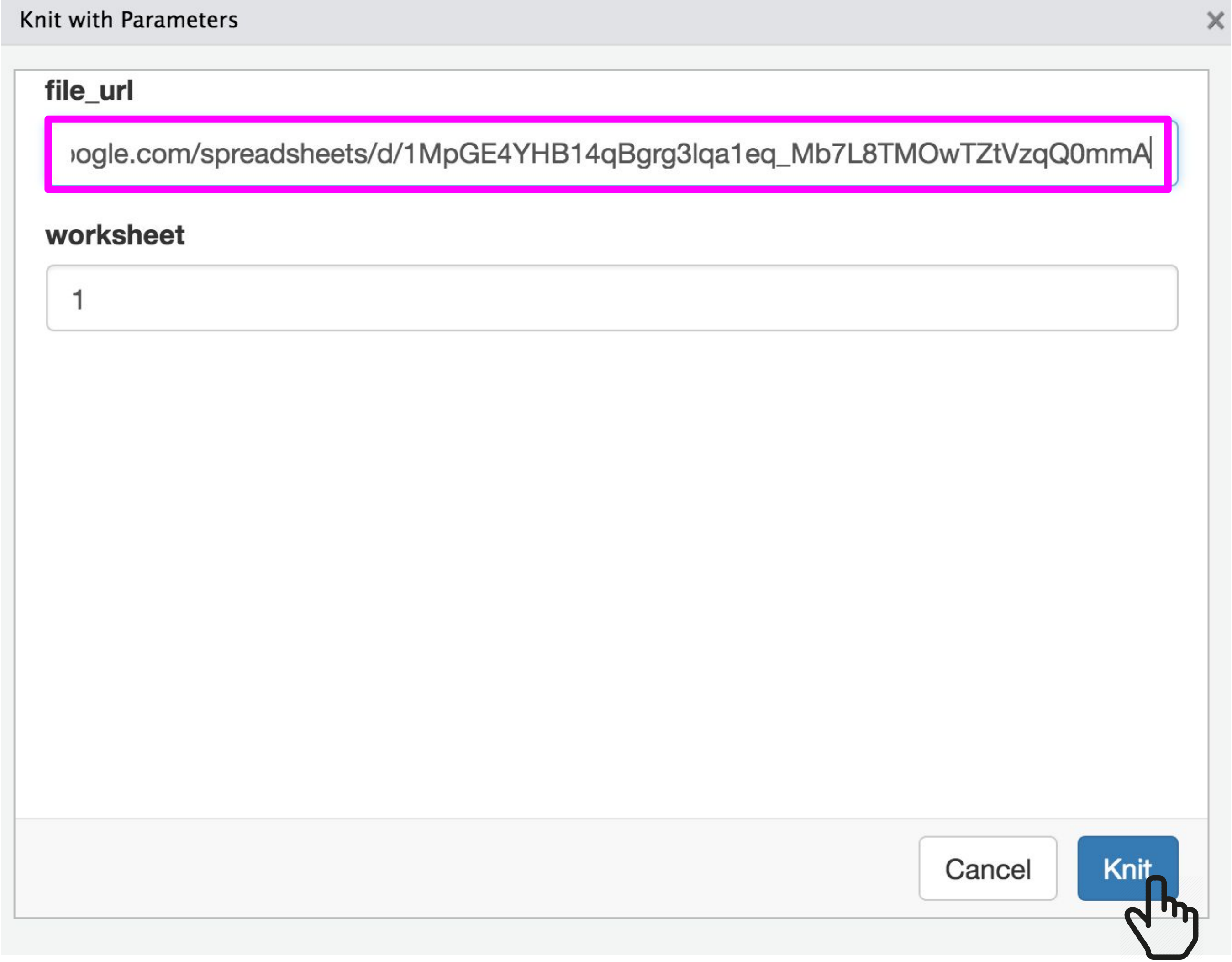


file_url

ask

worksheet

1



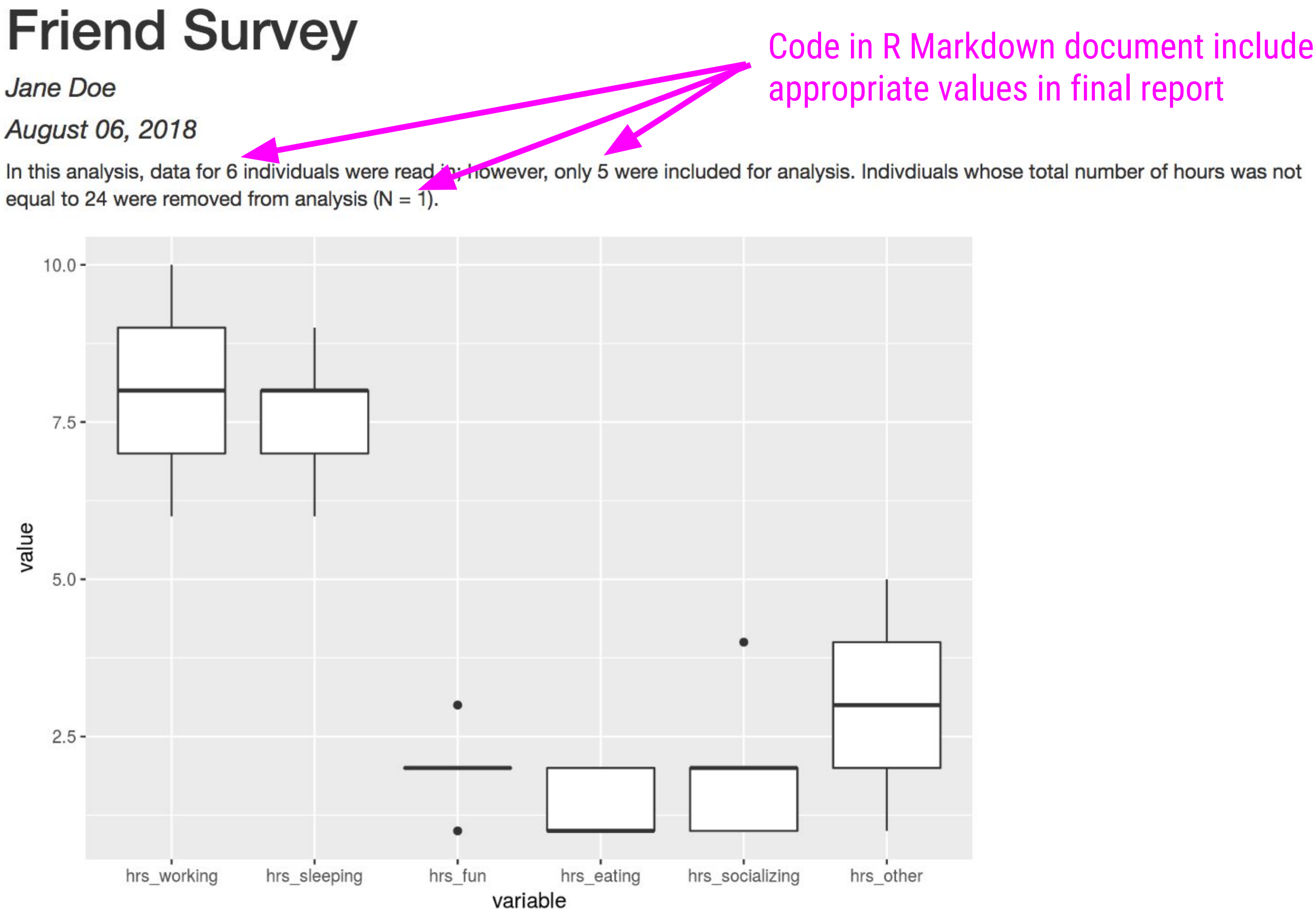
Knit with Parameters

file_url

ogle.com/spreadsheets/d/1MpGE4YHB14qBgrg3lqa1eq_Mb7L8TMOwTZtVzqQ0mmA

worksheet

1



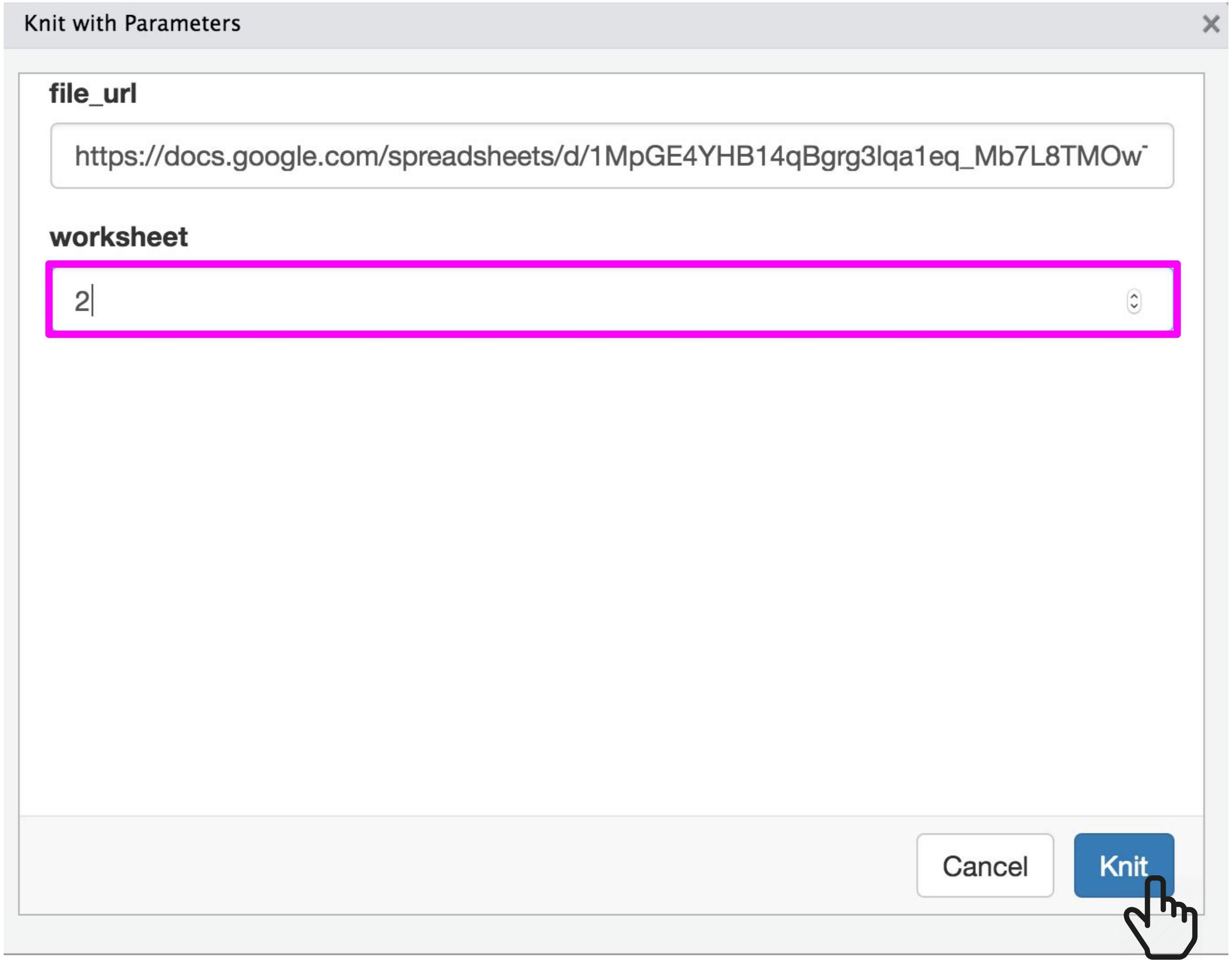
Friend Survey

Jane Doe

August 06, 2018

In this analysis, data for 6 individuals were read in; however, only 5 were included for analysis. Indivdiuals whose total number of hours was not equal to 24 were removed from analysis (N = 1).

Code in R Markdown document includes appropriate values in final report

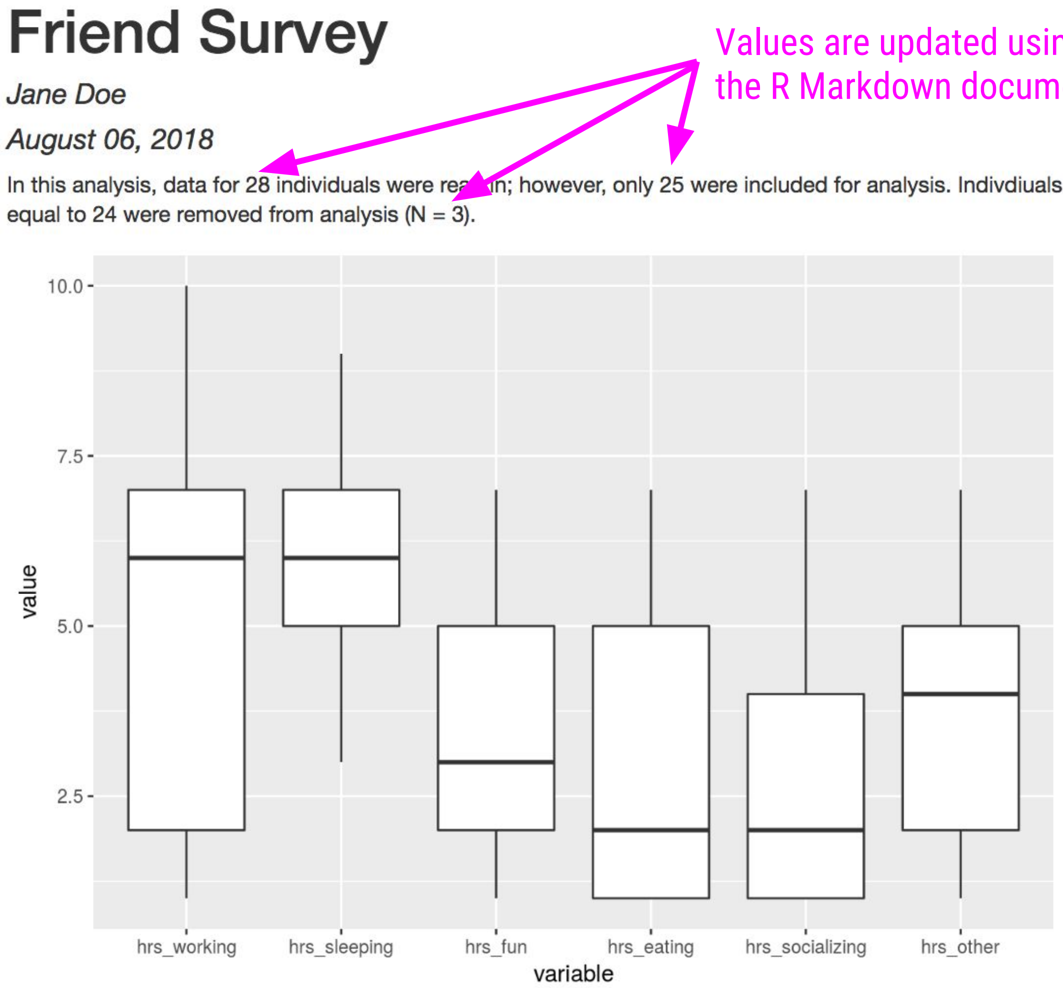


file_url

https://docs.google.com/spreadsheets/d/1MpGE4YHB14qBgrg3lqa1eq_Mb7L8TMOw

worksheet

2



Friend Survey

Jane Doe

August 06, 2018

In this analysis, data for 28 individuals were read in; however, only 25 were included for analysis. Indivdiuals whose total number of hours was not equal to 24 were removed from analysis (N = 3).

Values are updated using the code in the R Markdown document!

The same pipeline with updated data automatically updates the report!