HONOR *the* FUTURE

THE CAMPAIGN FOR THE UNIVERSITY OF VIRGINIA

READY, SET, BAKE

Recipes for Reproducible Reporting with RMarkdown

James Rogol | CASE Drive/ | 25 March 2020

re·pro·duc·i·bil·i·ty, *noun*

- Repeatable
- Identical
- Shareable

FOLLOW ALONG!

 github.com/jrogol/TBD

 **Reports** (HTML & PDF)

 **Markdown** (.Rmd)

CHALLENGE N° 1

THE SIGNATURE

CHALLENGE N° 1

THE SIGNATURE

1-TheSignature.pdf

Contact Report Summary

Month of August

Major Gifts

05 March 2020

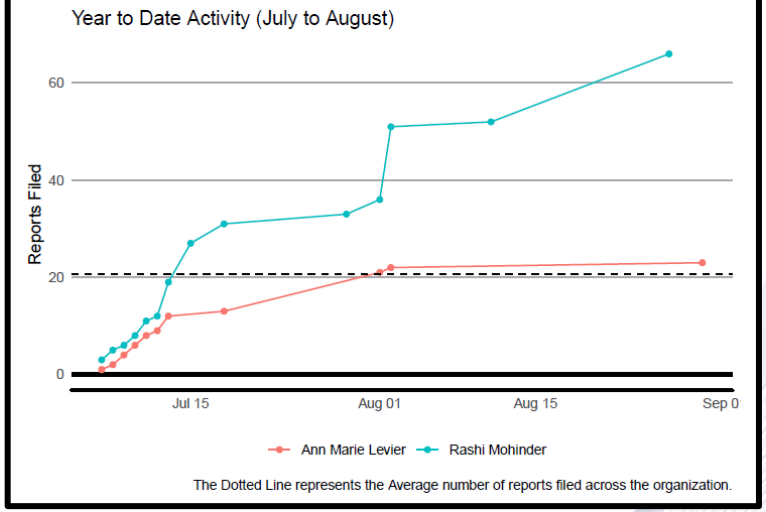
Reports filed in August, by Type.

	Email	Phone	Visit	Total
Ann Marie Levier	3	3	4	10
Rashi Mohinder	10	5	18	33
Total	13	8	22	43

Reach and Outcomes

	Reports	Unique Donors	Positive Outcomes	Negative Outcomes
Ann Marie Levier	23	21	0	23
Rashi Mohinder	66	59	16	50

Year to Date Summary



CHALLENGE N° 1

THE SIGNATURE


```yaml
---
title: "Contact Report Summary"
subtitle: "`r sprintf('Month of %s',params$endMonth)`"
author: "`r params$dept`"
date: "`r format(Sys.Date(),'%d %B %Y')`"
output:
  pdf_document:
params:
  endMonth: "August"
  dept: "Major Gifts"
---
```

Markdown/1-TheSignature.rmd

DATA AVAILABILITY

"If the first line of your #rstats script is

```
setwd("C:\Users\jenny\path\that\only\I\have")
```

I will come into your lab and
SET YOUR COMPUTER ON FIRE."

Jenny Bryan, Rstudio
Former Adjunct Professor
University of British Columbia

CHALLENGE N° 1

THE SIGNATURE

- Use `here`
- R code Chunk

```
```{r chunkName}

```
```

```
33 ```{r getData}
34 data <- read_csv(here("Data/contactReports.csv"))
35 ```
```

CHALLENGE N° 1

THE SIGNATURE

- Sections
- In-line code

  `` `r params$endMonth` ``

````
## Reports filed in `r params$endMonth`, by Type.

```{r byType}
deptData %>%
  filter(month == params$endMonth) %>%
  count(staffName, method) %>%
  spread(method,n) %>%
  adorn_totals() %>%
  adorn_totals("col") %>%
  rename(" " = staffName)  %>%
  kable(format = "latex")
```

## Reach and Outcomes

```{r byOutcome}
deptData %>%
  select(reportID,staffName, donor,outcome) %>%
  mutate(val = T) %>%
  spread(outcome, val, fill = F) %>%
  group_by(staffName) %>%
  summarize(reports = n(),
            uniqueDonors = n_distinct(donor),
            positive = sum(Positive),
            negative = sum(Negative)
            ) %>%
  kable(format = "latex",
        col.names = c("",
                      "Reports",
                      "Unique Donors",
                      "Positive Outcomes",
                      "Negative Outcomes"))
```
````

CHALLENGE N° 1

THE SIGNATURE

- Rendering
- Parameters

````
170
171 ```{r render, eval=F}
172 rmarkdown::render(here("Markdown/1-TheSignature.Rmd"),
173                   output_dir = here("Reports"))
174
175
176 rmarkdown::render(here("Markdown/1-TheSignature.Rmd"),
177                   output_dir = here("Reports"),
178                   output_file = "1-TheSignature-AG_Sept",
179                   params = list(endMonth = "September",
180                                 dept = "Planned Giving"))
181 ```
````

```
  |........
|   5%
   inline R code fragments

  |................
|  10%
label: setup (with options)
List of 1
 $ include: logi FALSE

  |........................
|  14%
  ordinary text without R code

  |..............................
|  19%
label: getData
  |....................................
|  24%
  ordinary text without R code

  |............................................
|  29%
label: cleanData1 (with options)
List of 1
```

CHALLENGE N° 1

THE SIGNATURE

Contact Report Summary

Month of August

Major Gifts

05 March 2020

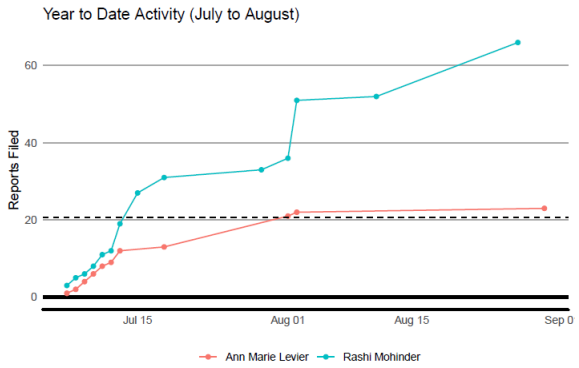
Reports filed in August, by Type.

	Email	Phone	Visit	Total
Ann Marie Levier	3	3	4	10
Rashi Mohinder	10	5	18	33
Total	13	8	22	43

Reach and Outcomes

	Reports	Unique Donors	Positive Outcomes	Negative Outcomes
Ann Marie Levier	23	21	0	23
Rashi Mohinder	66	59	16	50

Year to Date Summary



1-TheSignature.pdf

Contact Report Summary

Month of September

Planned Giving

05 March 2020

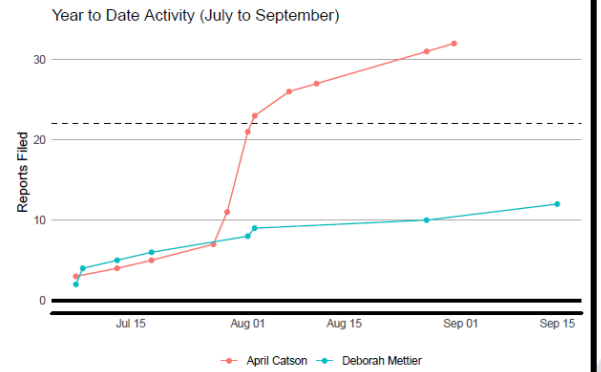
Reports filed in September, by Type.

	Visit	Total
Deborah Mettier	2	2
Total	2	2

Reach and Outcomes

	Reports	Unique Donors	Positive Outcomes	Negative Outcomes
April Catson	32	29	12	20
Deborah Mettier	12	12	4	8

Year to Date Summary



1-TheSignature-AG_sept.pdf

CHALLENGE N° 2

THE TECHNICAL


KOUIGN AMANN

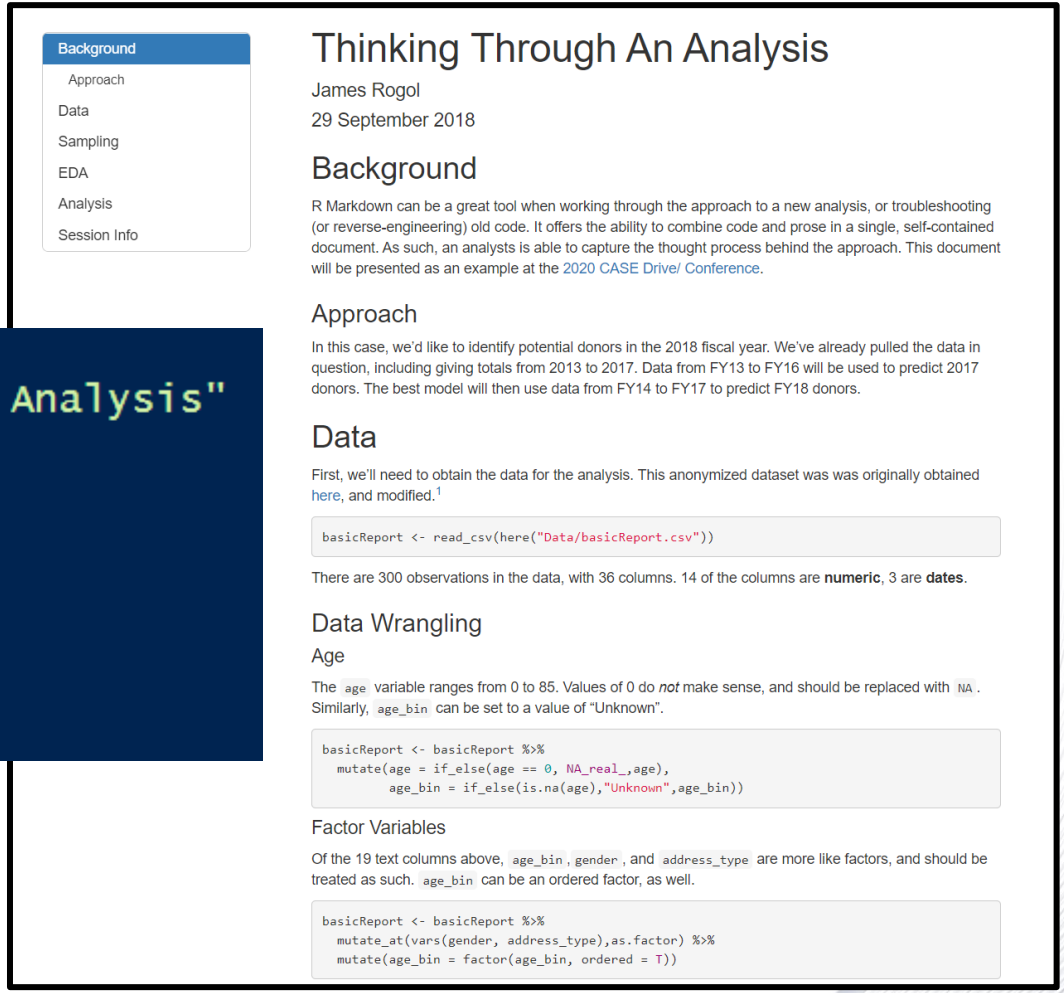

RASPBERRY
BLANCMANGE
with Langues du Chat

CHALLENGE N° 2

THE TECHNICAL

```yaml
---
title: "Thinking Through An Analysis"
author: "James Rogol"
date: "29 September 2018"
output:
  html_document:
    toc: true
    toc_float: true
---
```

1-TheTechnical.html
1-TheTechnical.md



Background
- Approach

Data
Sampling
EDA
Analysis
Session Info

Thinking Through An Analysis

James Rogol

29 September 2018

Background

R Markdown can be a great tool when working through the approach to a new analysis, or troubleshooting (or reverse-engineering) old code. It offers the ability to combine code and prose in a single, self-contained document. As such, an analysts is able to capture the thought process behind the approach. This document will be presented as an example at the 2020 CASE Drive/ Conference.

Approach

In this case, we'd like to identify potential donors in the 2018 fiscal year. We've already pulled the data in question, including giving totals from 2013 to 2017. Data from FY13 to FY16 will be used to predict 2017 donors. The best model will then use data from FY14 to FY17 to predict FY18 donors.

Data

First, we'll need to obtain the data for the analysis. This anonymized dataset was was originally obtained here, and modified.[1]

```
basicReport <- read_csv(here("Data/basicReport.csv"))
```

There are 300 observations in the data, with 36 columns. 14 of the columns are **numeric**, 3 are **dates**.

Data Wrangling

Age

The `age` variable ranges from 0 to 85. Values of 0 do *not* make sense, and should be replaced with `NA`. Similarly, `age_bin` can be set to a value of "Unknown".

```
basicReport <- basicReport %>%
  mutate(age = if_else(age == 0, NA_real_,age),
         age_bin = if_else(is.na(age),"Unknown",age_bin))
```

Factor Variables

Of the 19 text columns above, `age_bin`, `gender`, and `address_type` are more like factors, and should be treated as such. `age_bin` can be an ordered factor, as well.

```
basicReport <- basicReport %>%
  mutate_at(vars(gender, address_type),as.factor) %>%
  mutate(age_bin = factor(age_bin, ordered = T))
```

CHALLENGE N° 2

THE TECHNICAL

- Prose
  - Text Formatting
- Code Chunks
  - Named
  - Simple, monotonic

````
## Data

First, we'll need to obtain the data for the analysis. This anonymized dataset was was originally obtained [here](https://www.kaggle.com/michaelpawlus/fundraising-data), and modified.[^modify]

```{r getData}
basicReport <- read_csv(here("Data/basicReport.csv"))
```

There are `r nrow(basicReport)` observations in the data, with `r ncol(basicReport)` columns. `r ncol(select_if(basicReport,is.numeric))` of the columns are **numeric**, `r ncol(select_if(basicReport,~inherits(.,"POSIXct")))` are **dates**.


### Data Wrangling

#### Age

The `age` variable ranges from `r min(basicReport$age, na.rm = T)` to `r max(basicReport$age, na.rm = T)`. Values of 0 do _not_ make sense, and should be replaced with `NA`. Similarly, `age_bin` can be set to a value of "Unknown".

```{r cleanAge}
basicReport <- basicReport %>%
  mutate(age = if_else(age == 0, NA_real_,age),
         age_bin = if_else(is.na(age),"Unknown",age_bin))
```
````

CHALLENGE N° 2

THE TECHNICAL

- Randomness
- Functions
  - Simple, monotonic

Sampling

```
trainingPer <- .8
```

In order to measure model performance, we'll need to split the data into *Training* and *Testing* sets. 80% of the data will be used for training, and the other 20% will be used for testing.

```
set.seed(2020)

trainingInd <- sample(nrow(basicReport), nrow(basicReport)*trainingPer)

training <- basicReport[trainingInd,]

testing <- basicReport[-trainingInd,]
```

In the future, we can use the above as a model for a function. It will take three arguments: `df` (the data frame in question), `seed` for the random seed, and `per` for the percentage of observations to use as training.

```
sampleData <- function(df, per, seed = 2020){

  # Basic Error Handling
  stopifnot(inherits(df, "data.frame"))

  set.seed(seed)

  ind <- sample(nrow(df),nrow(df)*per)

  training <-  df[ind,]

  testing <- df[-ind,]

  out <- list()

  out$training <- training
  out$testing <- testing


  return(out)
}
```

CHALLENGE N° 2

THE TECHNICAL

- Randomness
- Functions
  - Simple, monotonic
- Dependencies

Session Info

Analysis and Report generated on Windows 10 x64 (build 17763) using R version 3.5.3 (2019-03-11).

```
# Results = 'asis' renders the output as if it were text in the markdown
doc. In
# this case, the script renders as a bulletted list.

# Get a list of attached packages
packages <- sessioninfo::package_info()

# Combine Package name and version number
p <- sprintf("%s (%s)",
             packages$package[packages$attached == T],
             packages$loadedversion[packages$attached == T])

# Iterate over the packages, adding a hypen and space at the start of th
e string
# (to create a bulletted list) and a new line character at the end of th
e string
# to list each package as a new bullet point.
for (x in p) {
  cat("- ", x, "\n")
}
```

- dplyr (0.8.3)
- forcats (0.4.0)
- ggplot2 (3.2.1)
- here (0.1)
- janitor (1.1.1)
- kableExtra (1.1.0)
- knitr (1.28)
- readr (1.3.1)
- tidyr (0.8.3)

CHALLENGE N° 3

THE SHOWSTOPPER

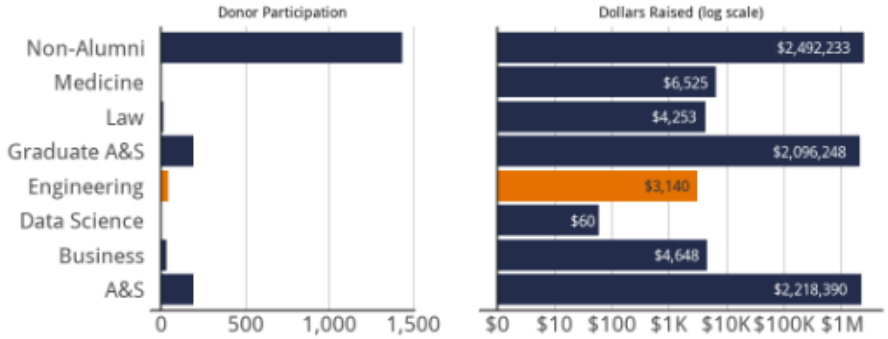
FY 2020 Giving Profile

School of Engineering Alumni

10 March 2020

Overview

43 Alumni from the School of Engineering gave in the 2020 Fiscal Year. This accounts for 2% of donor participation.



By Region

Of the 1,909 donors in FY 2020, 1,905 (100%) have a valid Zip Code.

418 entities reside outside of a CBSA. The table below shows the top 10 regions by count.

CHALLENGE N° 3

THE SHOWSTOPPER

- Custom CSS
  - JQuery Logo

UNIVERSITY of VIRGINIA

- Overview
- By Region
- Session Info

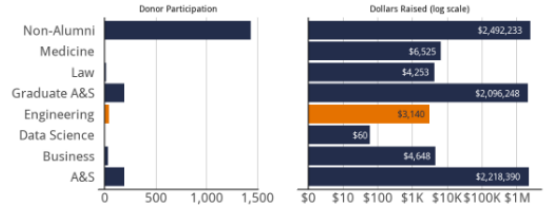
FY 2020 Giving Profile

School of Engineering Alumni

10 March 2020

Overview

43 Alumni from the School of Engineering gave in the 2020 Fiscal Year. This accounts for 2% of donor participation.



By Region

Of the 1,909 donors in FY 2020, 1,905 (100%) have a valid Zip Code.

418 entities reside outside of a CBSA. The table below shows the top 10 regions by count.

School of **Engineering** Alumni

Donor Participation in FY 2020

CBSA	Alumni	Total Donors	Alumni %
Portland-South Portland, ME	2	7	29%
Evanston, WY	1	1	100%
Heber, UT	1	1	100%
Ithaca, NY	1	1	100%
Monroe, MI	1	1	100%
Allentown-Bethlehem-Easton, PA-NJ	1	2	50%
Concord, NH	1	2	50%

CHALLENGE N° 3

THE SHOWSTOPPER

- Custom CSS
  - JQuery Logo
- SQL Chunks
  - Connection

```
con <- advancementtools::cx_Oracle(usr = "BI",
                                   pw =
key_get("advance","BI","adv"))

```

```
```{sql, output.var = "cx_geography"}
-- This SQL query creates a crosswalk between Zip codes and Core-Based
-- Statistical Areas (CBSA), Metropolitan Statistical Areas (MSA) and Combined
-- Statistical Areas (CSAs) via the 5-digit county-level FIPS code.

-- Note the "output.var" chunk argument, which saves the results as an R object.

WITH geo as (
    SELECT  FIP5
            ,title
            ,geography
    FROM    cx_geo
    UNPIVOT (title for geography in (
                                    county as 'County'
                                    ,cbsa_title     as 'CBSA'
                                    ,csa_title as 'CSA'
                                    ,msa_title as 'MSA'
                                    )
            )
    )


SELECT      DISTINCT ZIP            ZIPCODE
            ,title
FROM        cx_zip
INNER JOIN  geo
USING       (FIP5)
-- SQL chunks can use R variables using `?{variable_name}`
WHERE       geography = ?geoType
```
```

CHALLENGE N° 3

THE SHOWSTOPPER

- Custom CSS
  - JQuery Logo
- SQL Chunks
  - Connection

```
con <- advancementtools::cx_Oracle(usr = "BI",
                                   pw =
key_get("advance","BI","adv"))

```

```
knitr::opts_chunk$set(echo = FALSE,
                      message = F,
                      warning = F,
                      fig.showtext = TRUE,
                      # dpi = 72,
                      # The SQL connection needs to be passed to
any SQL chunks,
                      # and can be set here.
                      connection = "con")
```
```

```

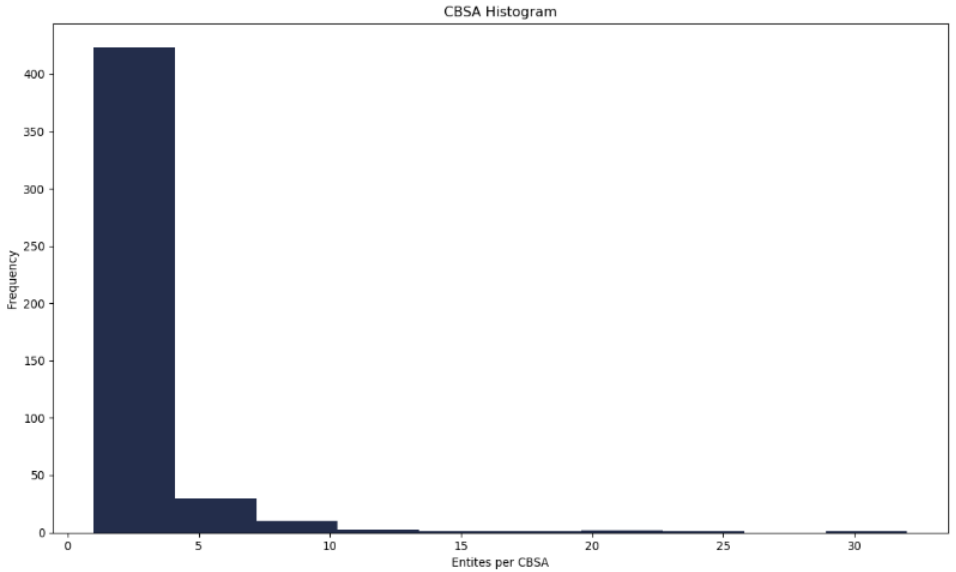
```
WITH geo as (
    SELECT  FIP5
            ,title
            ,geography
    FROM    cx_geo
    UNPIVOT (title for geography in (
                                    county as 'County'
                                    ,cbsa_title     as 'CBSA'
                                    ,csa_title as 'CSA'
                                    ,msa_title as 'MSA'
                                    )
            )
    )


SELECT      DISTINCT ZIP            ZIPCODE
            ,title
FROM        cx_zip
INNER JOIN  geo
USING       (FIP5)
-- SQL chunks can use R variables using `?{variable_name}`
WHERE       geography = ?geoType
```

CHALLENGE N° 3

THE SHOWSTOPPER

- Custom CSS
  - JQuery Logo
- SQL Chunks
  - Connection



CHALLENGE N° 3

THE SHOWSTOPPER

- Custom CSS
  - JQuery Logo
- SQL Chunks
  - Connection
- Python Chunks
  - Reticulate

````
```{python geoHist}
import pandas as pd
import numpy as np

np.random.seed(2020)

titles = r.combo['TITLE'].value_counts()

# Remove NA values
t2 = titles.drop('NA')

pl = t2.plot.hist(figsize = (14,8),
                  title = "%s Histogram" % (r.geoType),
                  color = '#232D4B')

pl.set_xlabel("Entites per %s" % (r.geoType))

# pl.show()
```

`r scales::comma(py$titles["NA"])` individuals do __not__ have a valid `r geoType`.

```{r geoTable}
geoTable <- table(py$titles)


maxRegions <- names(geoTable[geoTable == max(geoTable)])
```
````

CHALLENGE N° 3

THE SHOWSTOPPER

9 schools

6 years

54 reports

CHALLENGE N° 3

THE SHOWSTOPPER

```r
library(dplyr)

schools <- readr::read_csv(here::here("Data/donorBio.csv")) %>%
  # Only Alumni have a SCHOOL, replace missing values
  distinct(SCHOOL = tidyr::replace_na(SCHOOL,"Non-Alumni")) %>%
  pull()

years <- 2015:2020

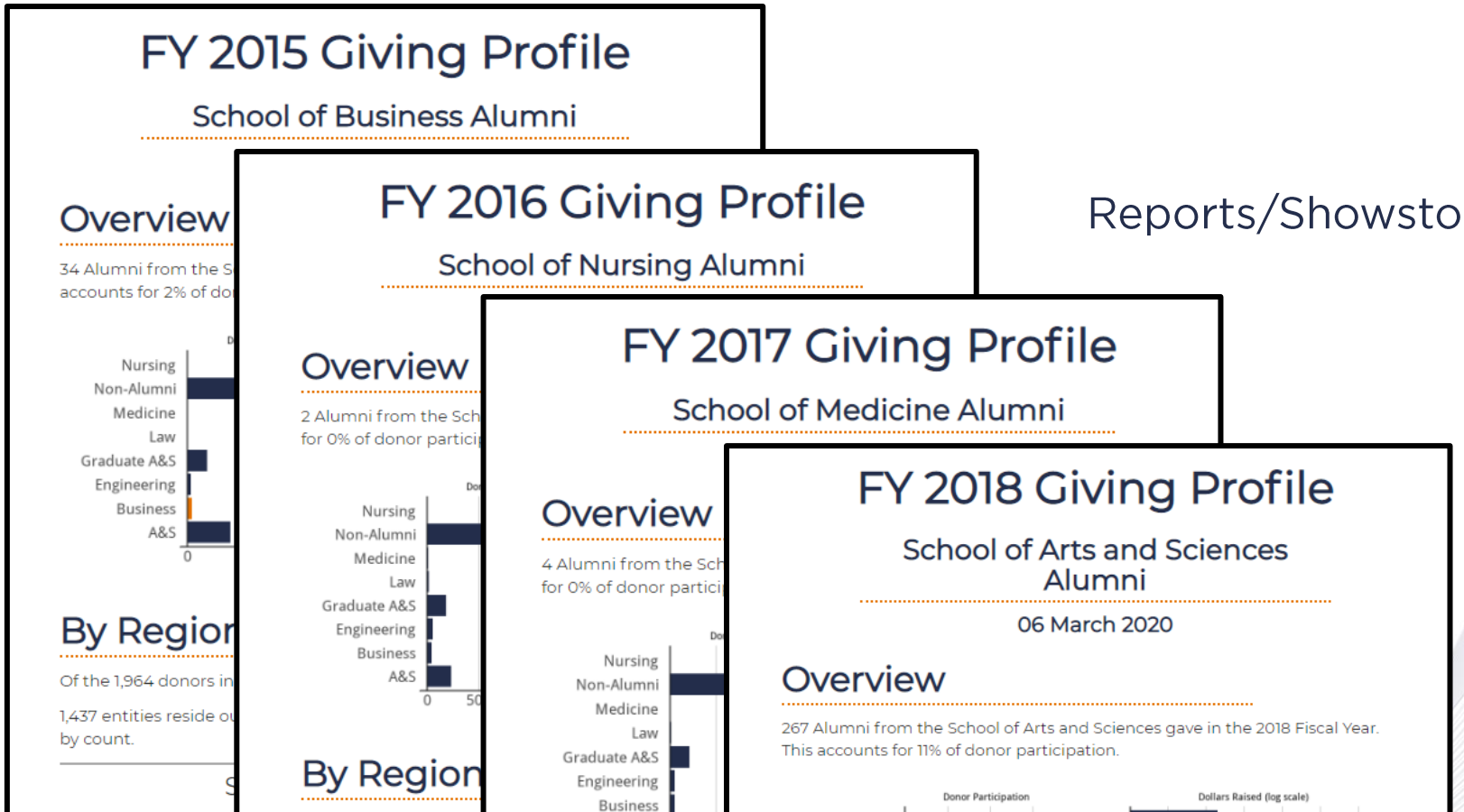

grid <- expand.grid(schools, years)

start <- Sys.time()
purrr::pwalk(grid,
             ~rmarkdown::render(here::here("Markdown/3-TheShowstopper.Rmd"),
             output_file = paste0(stringr::str_remove_all(..1,"[:space:]|[:punct:]"),
                                  "FY",..2,".html"),
             output_dir = here::here("Reports/Showstopper"),
             params = list(school = ..1,
                           year = ..2,
                           geoType = "MSA")))
```

R/parameterizedReports.R

CHALLENGE N° 3

THE SHOWSTOPPER



Reports/Showstopper

CHALLENGE N° 3

THE SHOWSTOPPER

CONTACT

 rogol@virginia.edu

 /in/jrogol

 /jrogol

 /jrogol

RESOURCES

- Yihui Xie, J. J. Allaire and Garrett Grolemund. *R Markdown: The Definitive Guide.*
- Yihui Xie. "R Markdown Cookbook." 27 February 2020.
- R Markdown Cheat Sheet. RStudio.
- R Markdown. RStudio.
- Emily Riederer. "R Markdown Driven Development: the Technical Appendix." 1 February 2020.
- Hadley Wickham. *R for Data Science.* "Chapter 27: R Markdown."
- Paul Hively "Reproducible Data Science." DRIVE/ 2019.

REFERENCES

Great British Baking Show Logo from Twin Cities PBS.

"The Great British Bake Off." Wikipedia.

Jenny Bryan. "Ode to the *here* package."

Kouign Amann recipe and picture by David Lebovitz.

Raspberry Blancmange recipe and picture from Prue Leith.

Icons made by Smashicons, Gregor Cresnar and Pixel Perfect from www.flaticon.com, licensed CC 3.0 BY.

Data adapted from Kaggle's "Fundraising Data", curated by Michael Pawlus.

Ashutosh Nandeshwar and Rodger Devine. *Data Science for Fundraising.*

Package Hex stickers from rmarkdown, bookdown, blogdown, pkgdown, pagedown and xaringan.