I’ve done an experiment regarding package downloads from CRAN (or the RStudio CRAN mirror at least) and now it’s time to share the results.

library(dplyr) library(ggplot2) library(purrr) library(dlstats) library(flextable)

flextable\_style <- function(x) { x %>%

flextable() %>%

bold(part = "header") %>% # bold header

bg(bg = "#D3D3D3", part = "header") %>% # puts gray background behind the header row

align\_nottext\_col(align = "center", header = TRUE, footer = TRUE)

%>% # center alignment autofit()

}

# Introduction

When the first version (0.0.1.0) of SwimmeR was released on CRAN in October of 2019 it had very few features – just a couple functions for formatting times and doing course conversions. So imagine my surprise when I checked the package download stats 6 months later and found out SwimmeR v0.0.1.0 had been downloaded over 1700 times. SwimmeR was, and is, a fairly niche package – there just aren’t that many people in the world interested in both swimming and R, so 1700 seemed like a lot. And remember, SwimmeR v0.0.1.0 had very few features.

SwimmeR has been out for about a year now, has many more features, and a little over 6500 downloads. That’s still not many in the grand scheme of things, but it’s many more than I ever expected. For comparison here’s how SwimmeR stacks up versus dplyr.

df <- cran\_stats(c("SwimmeR", "ThreeWiseMonkeys", "dplyr"))

df <- df %>% group\_by(package) %>%

mutate(days = cumsum(as.numeric(end - start, units = "days")), total\_downloads = cumsum(downloads),

package = as.character(package))

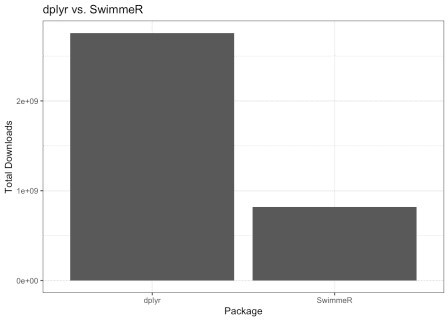
df %>%

filter(package != "ThreeWiseMonkeys") %>% group\_by(package) %>%

ggplot(aes(x = package, y = sum(downloads))) + geom\_col() +

labs(y = "Total Downloads", x = "Package",

title = "dplyr vs. SwimmeR") + theme\_bw()



# Too Many Downloads

Those initial 1700 downloads for a minimally featured SwimmeR v0.0.1.0 seemed so excessive that I got suspicious and decided to do an experiment. Allow me to introduce to the world the ThreeWiseMonkeys package. The documentation says “[ThreeWiseMonkeys d]oes nothing useful…”, which is mostly, but not quite, true. The reality is that ThreeWiseMonkeys is intended to do nothing useful for its users – the people who download it. I wrote ThreeWiseMonkeys in April 2020, released it to CRAN and told no one. What ThreeWiseMonkeys has been doing since then is steadily accruing downloads at a rate I’m prepared to call baseline. That’s what ThreeWiseMonkeys does – it marks out the CRAN downloads floor for me, the (supposedly) only person who knew or cared about it. It has about 2000 downloads as of this writing. Take a look versus SwimmeR:

df %>%

filter(package != "dplyr") %>% ggplot(aes(

x = end,

y = total\_downloads, group = package, color = package

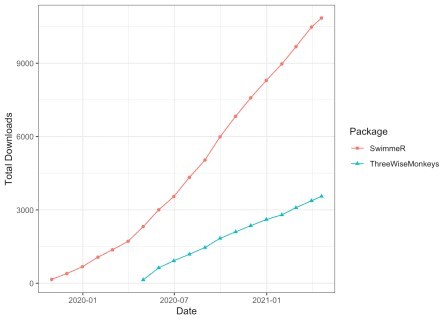
)) +

geom\_line() +

geom\_point(aes(shape = package)) + theme\_bw() +

labs(y = "Total Downloads", x = "Date",

color = "Package", shape = "Package")



Now if we line up SwimmeR and ThreeWiseMonkeys by the number of says since their releases we can see something interesting:

df %>%

filter(package != "dplyr") %>% ggplot(aes(

x = days,

y = total\_downloads, group = package, color = package

)) +

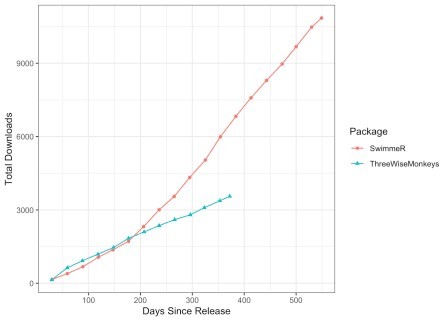
geom\_line() +

geom\_point(aes(shape = package)) + theme\_bw() +

labs(y = "Total Downloads",

x = "Days Since Release", color = "Package",

shape = "Package")



# SwimmeR Gets Some PR

What happened around April of 2020, 200 days after SwimmeR was released? SwimmeR and ThreeWiseMonkeys had pretty much identical download rates until then, but all of a sudden SwimmeR started trending upward much faster. Well, COVID happened, and maybe folks took to R rather than socializing. That’s possible but I’m going to set it aside for the moment, because it doesn’t explain why SwimmeR jumped but ThreeWiseMonkeys didn’t. As far as SwimmeR specific changes in that time range SwimmeR v0.2.0 was released on April 10th, with many more features, effectively beginning its life as a useful package. Let’s take a look at the difference in downloads, with the post April 2020 SwimmeR package renamed as "SwimmeR + Blog".

df <- df %>%

filter(package != "dplyr") %>% # don't need dplyr stats any more mutate(package = case\_when(

package == "SwimmeR" &

start > as.Date("2020-04-01") ~ "SwimmeR + Blog", # rename package for month of April

TRUE ~ package

))

ggplot(data = df, aes(group = package, color = package)) + geom\_point(aes(

x = days,

y = total\_downloads

)) +

geom\_smooth( # fit lines for SwimmeR and ThreeWiseMonkeys (want full range)

data = df %>%

filter(package != "SwimmeR + Blog"), aes(x = days,

y = total\_downloads,

group = package), method = "lm", fullrange = TRUE, se = FALSE

) +

geom\_smooth( # fit line for SwimmeR + Blog (don't want full range on this one)

data = df %>%

filter(package == "SwimmeR + Blog"), aes(x = days,

y = total\_downloads, group = package), method = "lm", fullrange = FALSE, se = FALSE

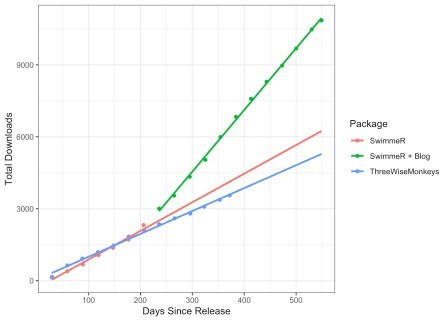
) +

scale\_x\_continuous(breaks = seq(0, 400, 25)) +

scale\_y\_continuous(breaks = seq(0, 7000, 1000)) + labs(y = "Total Downloads",

x = "Days Since Release", color = "Package") +

theme\_bw()



What this means is that of the 6500 downloads SwimmeR has as of October 28th, 2020, about 4000 can be provisionally attributed to whatever this CRAN baseline is. The other ~2500 are more likely to be the result of actual people who share my interests in swimming and R – thank you friends, I hope SwimmeR is helping you! Now lets get the slope (total downloads per day) for each package and compare download rates for SwimmeR

df %>%

group\_split() %>% # breaks into separate dataframes for each group (package)

map(~lm(total\_downloads ~ days, data = .x)) %>% # apply lm function to each dataframe

map\_df(broom::tidy) %>% # clean up results by converting back to single dataframe - broom heh heh

filter(term == "days") %>% # only want the slope term mutate(package = unique(df$package)) %>% # add package names back in select(package, "download rate" = estimate) %>%

mutate(slope = round(`download rate`, 2)) %>% arrange(desc(slope)) %>%

flextable\_style()

|  |  |  |
| --- | --- | --- |
| package | download rate | slope |
| SwimmeR + Blog | 25.85800 | 25.86 |
| SwimmeR | 11.77588 | 11.78 |
| ThreeWiseMonkeys | 10.55745 | 10.56 |

The download rate for SwimmeR more than doubled (as of October 28th, 2020) after I started writing these articles and building out the package features – exactly what I was looking for! What’s the deal with ThreeWiseMonkeys and its 2000 downloads though?

# The Remaining Mystery

The dlstats package I used to collect download information queries the RStudio CRAN mirror. RStudio’s mirror is the default for RStudio users when they download packages, but if anything the numbers from dlstats are an undercount for total package downloads. There are other mirrors, and other development environments for R and people can download packages using them rather than RStudio. That still leaves some questions.

1. Who are these 2000+ people who downloaded ThreeWiseMonkeys and are they the same as the 1700+ people who downloaded SwimmeR v0.0.1.0?
2. How did they even know ThreeWiseMonkeys existed? I know that CRAN lists it and all, but how would people know to look?
3. Why did they download ThreeWiseMonkeys even having found it? It’s useless!

It seems possible that there’s some kind of maintenance/testing/checks that goes on in the background at CRAN, but I’m not aware of anything that would require the same (useless) package to be re-downloaded 10 times a day, every day, for 6 months. If you have any insight into what’s going on leave a comment or send me an email. I’m interested to know… Also don’t go and download ThreeWiseMonkeys – that’s not what it’s about. It’s useless – trust me.