Harvardx / MITx Courses Year 1

John Montroy
NYC Data Science Academy
Project 1

Introduction

- 2012 - 2013: HarvardX and MITx launches courses through edX platform
- Sampling of available courses:
  - The Ancient Greek Hero (HarvardX)
  - Health in Numbers: Quantitative Methods in Clinical & Public Health Research (HarvardX)
  - Introduction to Solid State Chemistry (MITx)
  - The Challenges of Global Policy (MITx)
- Totals:
  - 17 courses across 3 “semesters”
  - 597,692 unique users
  - 43,196 certificates of completion issued

Motivations + Objectives

Motivating Questions

- Poor completion rates - what to do?
  - Who's performing better? Worse?
- Targeted analysis
  - Identifying and describing groups of users
  - Testing efficacy of retainment campaigns

Objectives

- Focus Areas:
  - User Engagement
  - User Outcomes
- Coding Asides
  - developing re-usable ways to clean/aggregate data

Data

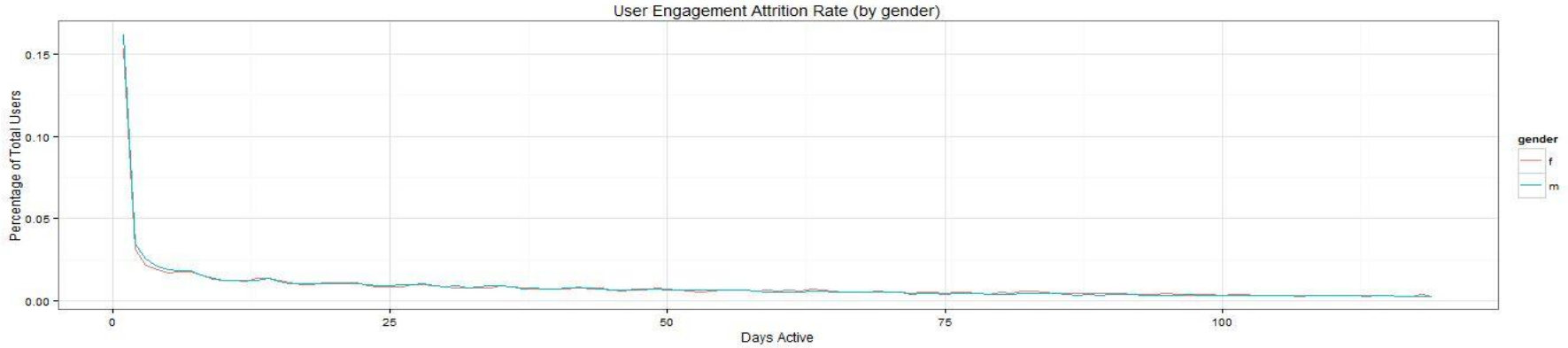
- HMXPC13_DI_v2_5-14-14.csv
  - 641,138 observations spanning 20 variables
  - 66.9 MB - not particularly big
  - Kinds of columns:
    - User-provided (age, year of birth, etc -- plenty of missingness!)
    - Administrative (number of forum posts, days active, final grade)
- Person Course Documentation.pdf
  - Business/technical logic behind dataset
  - Very nice to have!

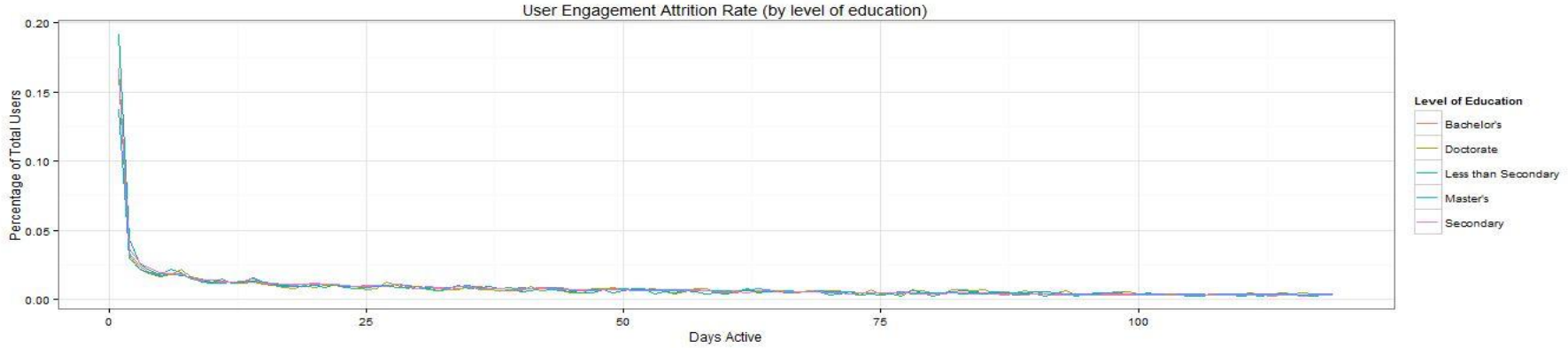
Coding Aside #1

```
# change all blanks in factor columns to NA
# edx[edx$YoB == "",] <- edx %>% filter(YoB == "") %>% mutate(YoB = NA)
edx <- edx %>% mutate(YoB = ifelse(YoB == "", NA, as.character(YoB)))
edx <- edx %>% mutate(gender = ifelse(gender == "", NA, as.character(gender)))
edx <- edx %>% mutate(LoE_DI = ifelse(LoE_DI == "", NA, as.character(LoE_DI)))
edx <- edx %>% mutate(start_time_DI = ifelse(start_time_DI == "", NA, as.character(start_time_DI)))
edx <- edx %>% mutate(last_event_DI = ifelse(last_event_DI == "", NA, as.character(last_event_DI)))
edx <- edx %>% mutate(final_cc_cname_DI = ifelse(final_cc_cname_DI == "", NA, as.character(final_cc_cname_DI)))
```

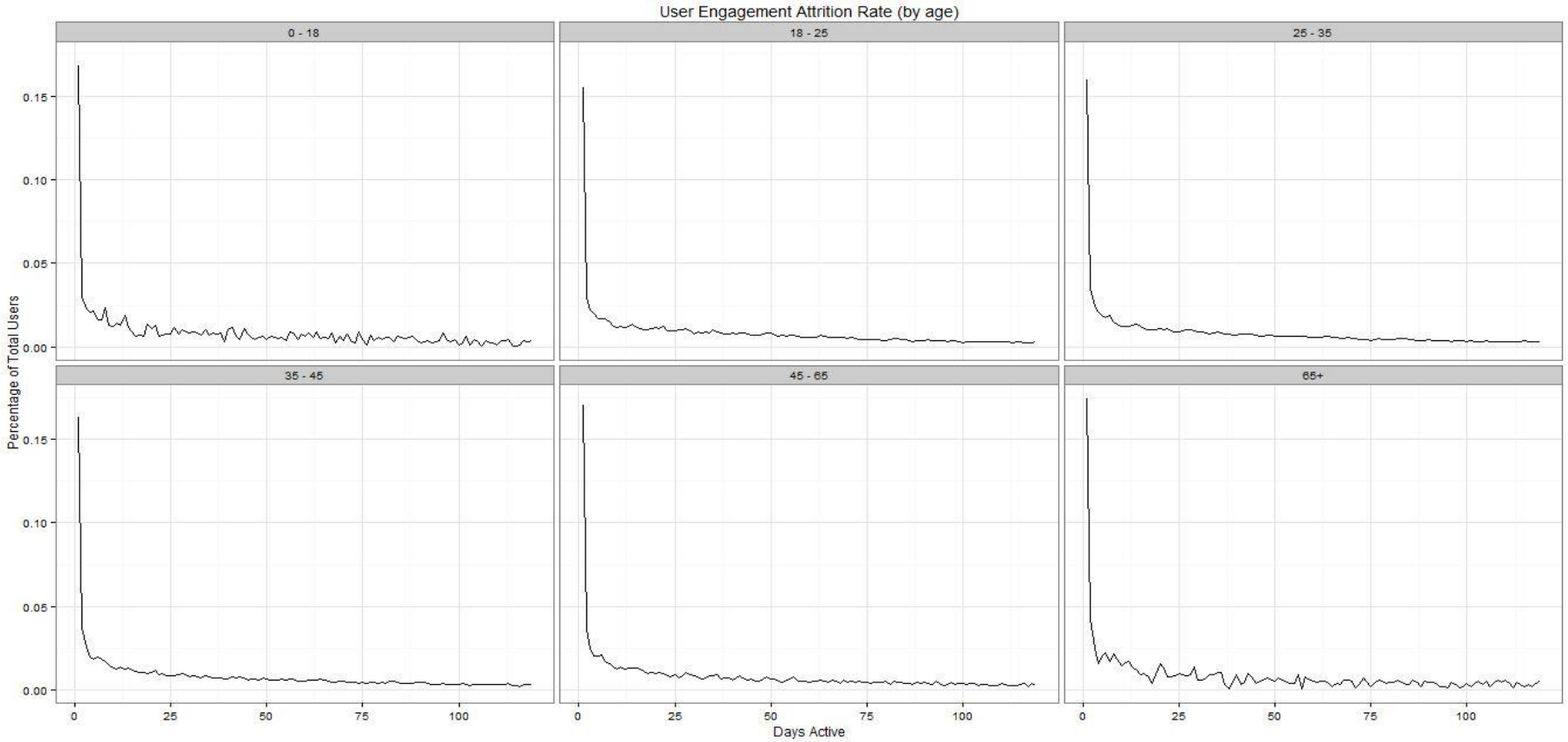
Silly! Let's make a function?

User Engagement (Gender, Education)





User Engagement (Age)

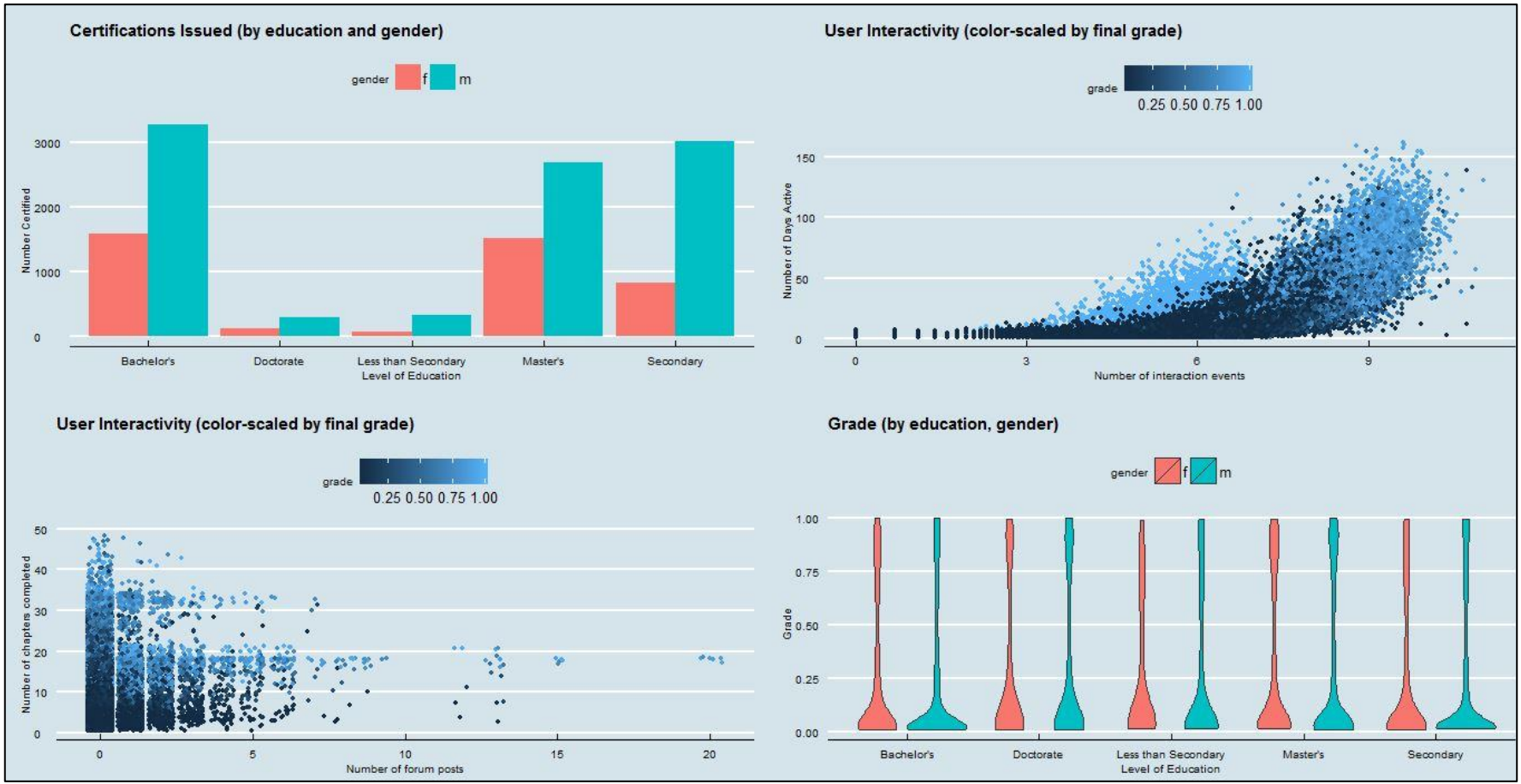


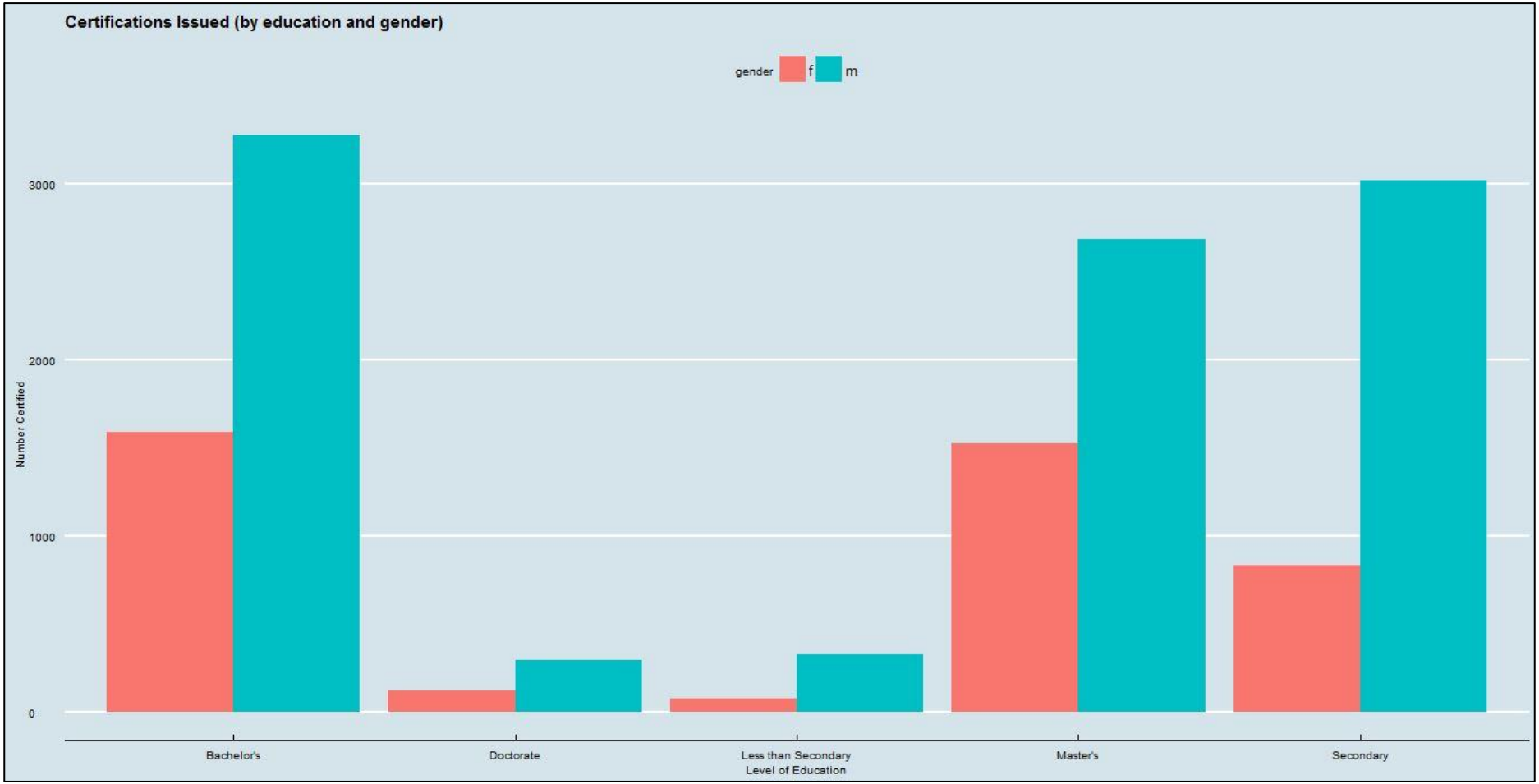
Coding Aside #2

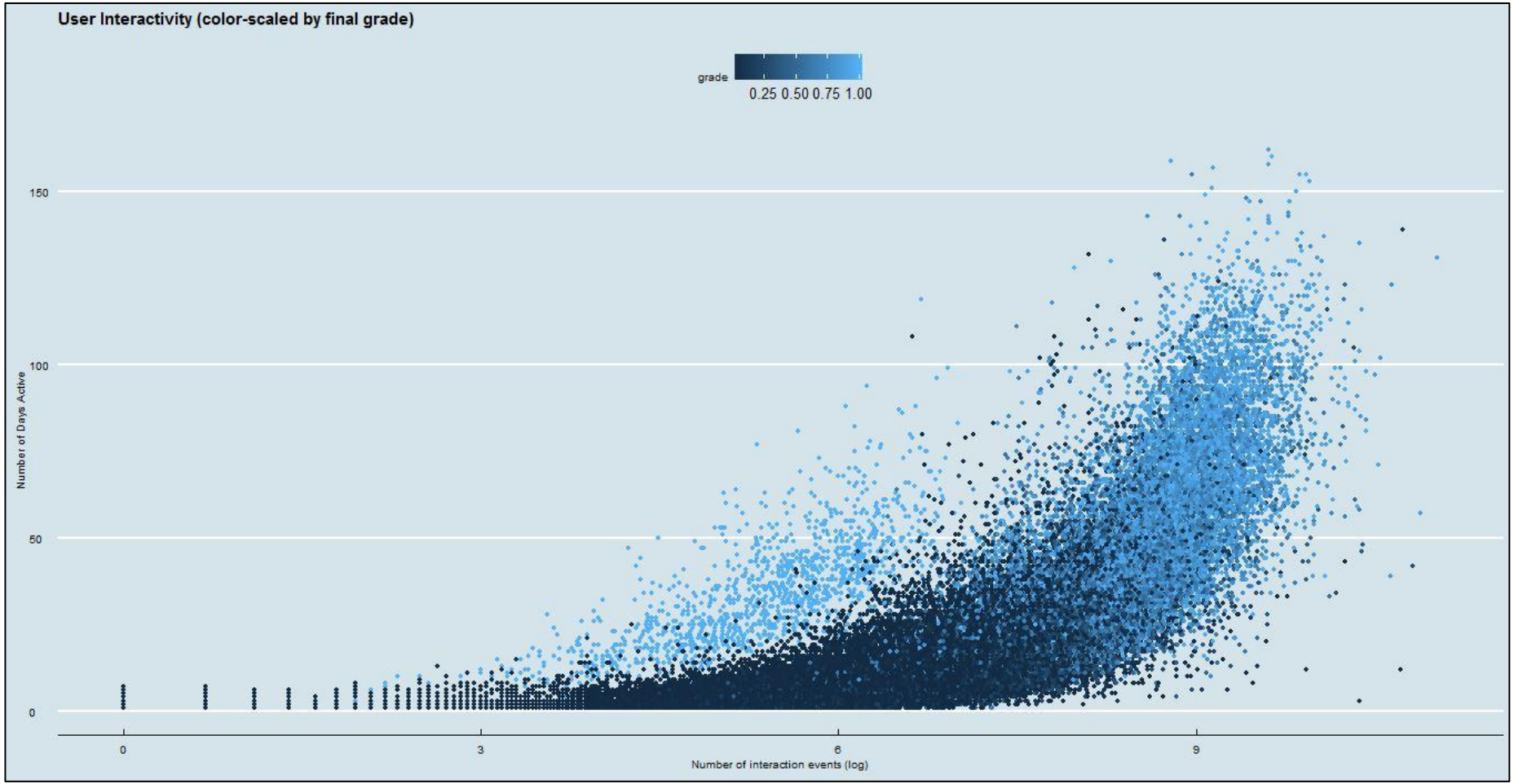
```r
# by education level
edx.edlevel <- edx %>%
  filter(active_length > 0, active_length < 120) %>%
  filter(!is.na(LoE_DI)) %>%
  group_by(LoE_DI, active_length) %>%
  summarise(count = n())

edx.edlevel.tots <- edx %>%
  filter(active_length > 0, active_length < 120) %>%
  filter(!is.na(LoE_DI)) %>%
  group_by(LoE_DI) %>%
  summarise(totcount = n())

edx.edlevel <- edx.edlevel %>% inner_join(edx.edlevel.tots, by = c("LoE_DI"))
edx.edlevel <- edx.edlevel %>% mutate(perc = count / totcount)
```

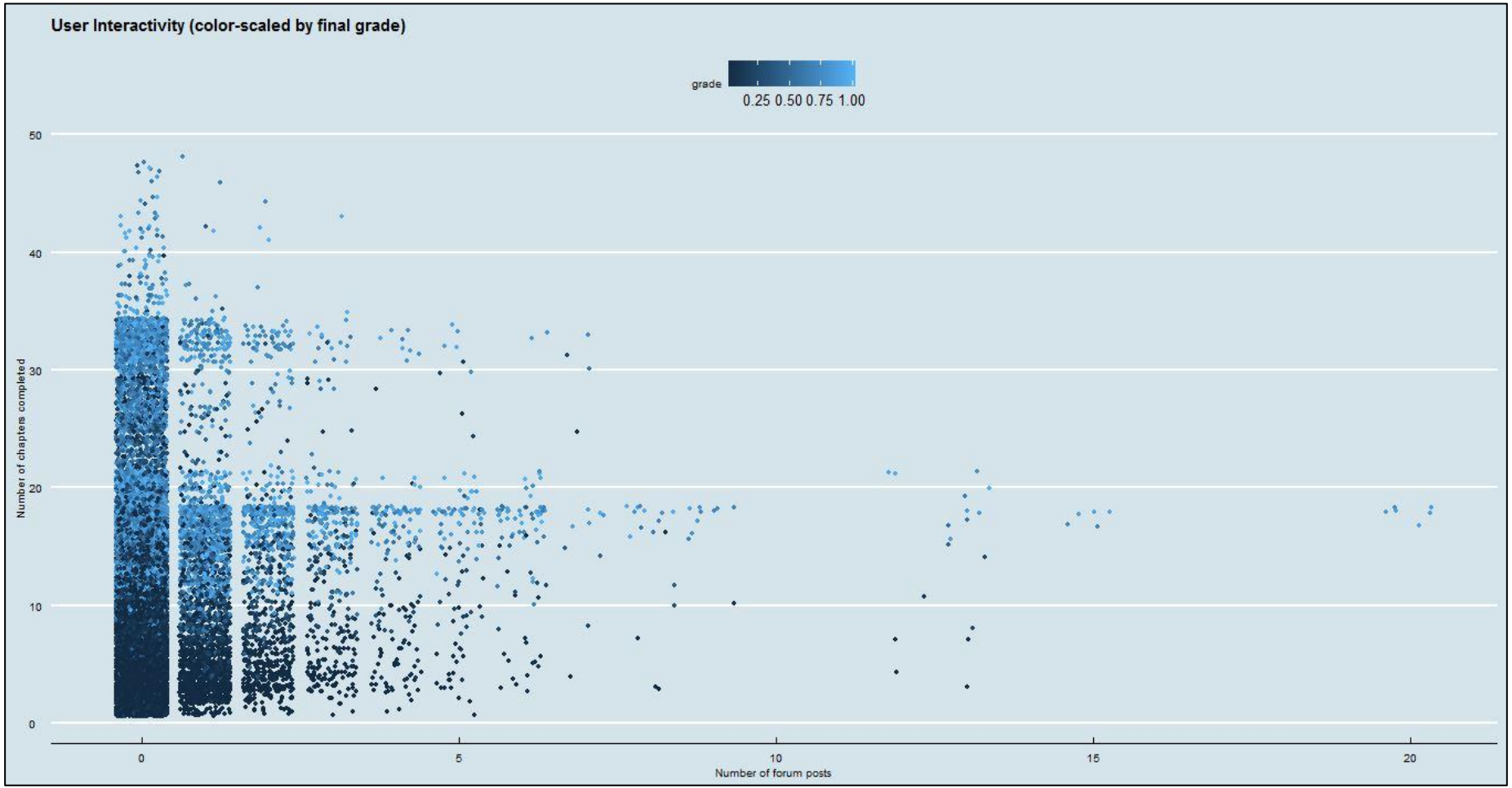
Repetitive! Let’s make a function?

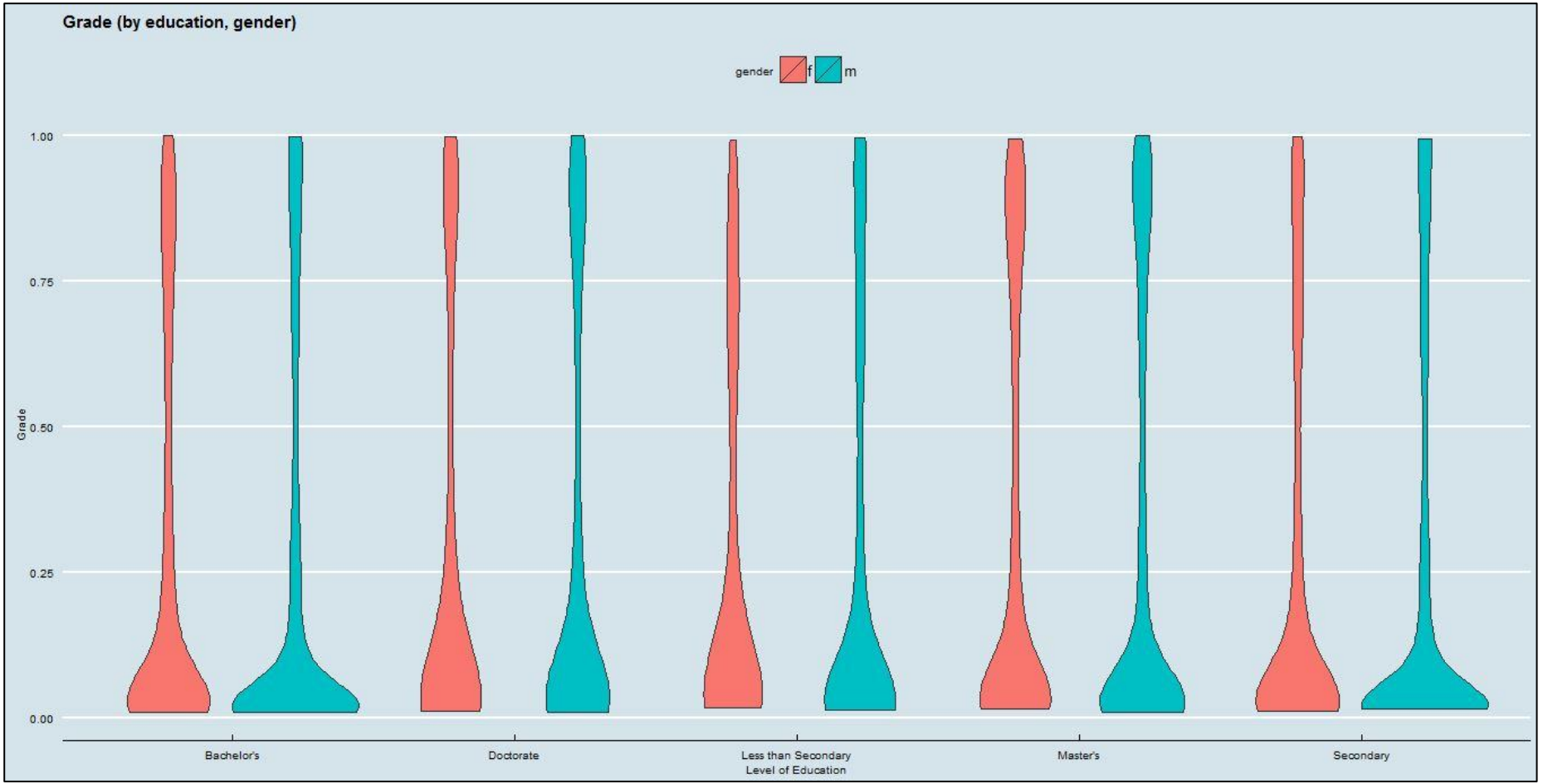
User Outcomes




Certifications Issued (by education and gender)
gender
f
m
3000
2000
1000
0
Number Certified
Bachelor's
Doctorate
Less than Secondary
Master's
Secondary
Level of Education


User Interactivity (color-scaled by final grade)
grade
0.25 0.50 0.75 1.00
150
100
50
0
Number of Days Active
0
3
6
9
Number of interaction events (log)


User Interactivity (color-scaled by final grade)
grade
0.25 0.50 0.75 1.00
50
40
30
20
10
0
Number of chapters completed
0
5
10
15
20
Number of forum posts


Grade (by education, gender)
gender
f
m
1.00
0.75
0.50
0.25
0.00
Grade
Bachelor's
Doctorate
Less than Secondary
Master's
Secondary
Level of Education

Next Steps + Lessons

- Email engagement campaign
  - Identify disengaged users
  - Set up experimental groups
  - Compare efficacy
- User classification
  - Identify and define subsets
  - Engage differently to leverage group strengths (“overachievers - show off!”)
- Personally - clever isn’t cute.