Clean Highest Degree

Aaron R. Williams – Urban Institute

2023-06-06

Introduction

This script cleans responses from a question about the highest degree attained for non-expired paying members of the American Economic Association, who opted into both the Directory of Members and receiving surveys sent by the association.

Data were extracted from internal systems. Members complete the "Directory of Members" form in order to be included in the directory, which used to be like a telephone book in the old days. They can choose to receive surveys sent out by the AEA.

The data contain 3,813 responses for the highest degree attained question. It reflects degree most recently completed as recorded in the directory, which should almost always yield the highest, but in (rare or maybe non-existing) cases where someone switches fields of study, is not guaranteed. The "degree" field is free text and contains all manor of notation. Free text was extracted, and normalized by code present in this script.

Setup

```
library(tidyverse)
library(here)
library(readxl)
```

```
highest_degree <- read_excel(
  here::here("data", "raw", "memberData.xlsx"),
  sheet = "Highest Degree",
  col_names = c("degree", "n")
)
```

Clean Highest Degree

The field for higest degree was open response and yielded an inconsistent array of responses. We clean up the responses with a few different steps:

1. Text preprocessing
    a. Convert all text to lower case
    b. Remove symbols, delete common prepositions, and remove the word economics.
2. Manually clean clear cases
3. Identify students without PhDs
4. Simplify degrees
5. Clean up international degrees
6. Make decisions about tough cases
7. Coarsen to groups of interest

```
#' Simplify a text response that matches a pattern
#'
#' @param string The vector containing responses
#' @param pattern A regular expression to match strings to change
#' @param replacement A new response
#'
#' @return A vector with cleaned responses
#'
simplify_response <- function(string, pattern, replacement) {

  print(paste("Observations changed:", sum(str_detect(string = string, pattern = pattern))

  if_else(
    condition = str_detect(string = string, pattern = pattern),
    true = replacement,
    false = string
  )

}
```

1. Text preprocessing

First, we simplify text by converting everything to lower case, removing special characters, dropping common prepositions, and dropping “econ” and “economics”.

```r
highest_degree <- highest_degree %>%
  mutate(
    degree_clean = degree,
    degree_clean = str_to_lower(degree_clean),
    degree_clean = str_remove_all(degree_clean, pattern = "\\."),
    degree_clean = str_remove_all(degree_clean, pattern = ","),
    degree_clean = str_remove_all(degree_clean, pattern = "-"),
    degree_clean = str_remove_all(degree_clean, pattern = "\\("),
    degree_clean = str_remove_all(degree_clean, pattern = "\\)"),
    degree_clean = str_remove_all(degree_clean, pattern = "\\["),
    degree_clean = str_remove_all(degree_clean, pattern = "\\]"),
    degree_clean = str_remove_all(degree_clean, pattern = "\\&"),
    degree_clean = str_remove_all(degree_clean, pattern = "•"),
    degree_clean = str_replace_all(degree_clean, pattern = " in ", replace = " "),
    degree_clean = str_replace_all(degree_clean, pattern = " of ", replace = " "),
    degree_clean = str_remove_all(degree_clean, pattern = "'"),
    degree_clean = str_remove_all(degree_clean, pattern = "economics"),
    degree_clean = str_remove_all(degree_clean, pattern = "econ"),
    degree_clean = str_squish(degree_clean)
  )
```

2. Manually clean clear cases

"ph d" is a very common response after preprocessing. We clean up several cases in this vein.

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "ph d", "phd"))
```

```
[1] "Observations changed: 6"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "d phil", "phd"))
```

```
[1] "Observations changed: 2"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "m b a", "ma"))
```

```
[1] "Observations changed: 1"
```

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "m sc", "ma"))
```

```
[1] "Observations changed: 3"
```

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "b sc", "ba"))
```

```
[1] "Observations changed: 1"
```

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "m ec", "ba"))
```

```
[1] "Observations changed: 1"
```

3. Identify students without PhDs

Next, we label students. This will avoid situations later where we might accidentally label "phd abd" or "phd student" as people with PhDs. We assume that anyone who is a candidate, ABD, or PhD student has an MA.

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "phd student", "ma"))
```

```
[1] "Observations changed: 3"
```

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "phd sought", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "phd ongoing", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "in process", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "anticipated", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "cand", "ma"))
```

```
[1] "Observations changed: 11"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "abd", "ma"))
```

```
[1] "Observations changed: 6"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "all but dissertation", "ma"))
```

```
[1] "Observations changed: 1"
```

4. Simplify degrees

The responses contain tons of extraneous information.

We work from the highest level degree (PhD or DBA) to the lowest level degree. This means if someone has a PhD and JD, their highest degree will be a PhD. If someone lists “ba and ma”, their highest degree will be an ma.

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "phd|^doctor|^dr|dphil|dba|edd|doc
```

```
[1] "Observations changed: 267"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "jd", "jd"))
```

```
[1] "Observations changed: 3"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "judicial expert|juris|llb", "jd")
```

```
[1] "Observations changed: 3"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "postdoc|post doc|pgd", "phd"))
```

```
[1] "Observations changed: 6"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "^master|mba|^ma|^ms|mphil|mpa|^am
```

```
[1] "Observations changed: 230"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "^ba|^bs|^bm|^ab|bphil|^sb|^bba|un
```

```
[1] "Observations changed: 68"
```

5. Clean up international degrees

We clean up a few international examples.

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "habil|hab", "phd"))
```

```
[1] "Observations changed: 8"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "docent", "phd"))
```

```
[1] "Observations changed: 2"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "docteur", "phd"))
```

```
[1] "Observations changed: 1"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "univdoz", "phd"))
```

```
[1] "Observations changed: 1"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "lhd", "phd"))
```

```
[1] "Observations changed: 1"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "mres", "ma"))
```

```
[1] "Observations changed: 3"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "volkswirt", "ma"))
```

```
[1] "Observations changed: 3"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "dea", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "licen", "ma"))
```

```
[1] "Observations changed: 3"
```

```
  # germany
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "mtech", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "diplomoekonom", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  # england
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "mlitt", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  # india
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "pgsem", "ma"))
```

```
[1] "Observations changed: 1"
```

```r
# spain
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "laurea", "ba"))
```

```
[1] "Observations changed: 2"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "sarjana ekonomi", "ba"))
```

```
[1] "Observations changed: 1"
```

```r
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "bacharel", "ba"))
```

```
[1] "Observations changed: 1"
```

Here are the results before we start working with edge cases.

```r
highest_degree %>%
  count(degree_clean, sort = TRUE, wt = n) %>%
  print(n = Inf)
```

```
# A tibble: 93 x 2
   degree_clean                                                          n
   <chr>                                                             <dbl>
 1 "phd"                                                              3059
 2 "ma"                                                                506
 3 "ba"                                                                117
 4 ""                                                                   14
 5 "jd"                                                                 14
 6 "professor"                                                           7
 7 "certificate"                                                         3
 8 "public policy"                                                       3
 9 "assistant professor"                                                 2
10 "diploma"                                                             2
11 "md"                                                                  2
12 "mpp"                                                                 2
```

13	"prof"	2
14	"acs"	1
15	"agrégation"	1
16	"agrégation des universités"	1
17	"alm finance"	1
18	"and finance"	1
19	"and finance aging"	1
20	"and japanese language literature"	1
21	"applied nd"	1
22	"applied ometrics"	1
23	"as business administration management"	1
24	"assistant research fellow"	1
25	"associate professor"	1
26	"b"	1
27	"be"	1
28	"business finance"	1
29	"business management"	1
30	"cert art intelligence business applications"	1
31	"certificate board leadership"	1
32	"certified business omist"	1
33	"cfa"	1
34	"cfa charterholder"	1
35	"computer science"	1
36	"continued education course"	1
37	"degree"	1
38	"dimploma hotel and tourism management"	1
39	"diploma postgrau en omia i gestiã³ de la hisenda autonã²mica i local"	1
40	"education specialist special educ credential eds"	1
41	"educator certificate school guidance"	1
42	"environmental and policy"	1
43	"executive education evidence policy design"	1
44	"fellowship"	1
45	"finance"	1
46	"financial"	1
47	"fiscal crimes expert commercial money laundering and terrorism financ~	1
48	"foreign affairs"	1
49	"frm"	1
50	"full professor scientific title"	1
51	"global studies ba"	1
52	"government and politics"	1
53	"graduate certificate intermediate survey methodology"	1
54	"graduate certificate ip law"	1
55	"graduate certificate public health"	1

56	"graduate degree and statistics"	1
57	"graduate diploma finance"	1
58	"graduate statistician gstat"	1
59	"graduated"	1
60	"honoured professor"	1
61	"international and finance"	1
62	"international applied specialist"	1
63	"llm"	1
64	"m pub pol mgt"	1
65	"m s"	1
66	"mbs"	1
67	"mdiv"	1
68	"me"	1
69	"med educational leadership policy and human development"	1
70	"methodology complex sales"	1
71	"mfa"	1
72	"mla mms med"	1
73	"mls"	1
74	"mmh"	1
75	"mom logistics and supply chain"	1
76	"mph"	1
77	"murp omic development"	1
78	"northwestern university"	1
79	"omic aanalysis specialist"	1
80	"omic sciences"	1
81	"organizational communication"	1
82	"p"	1
83	"post graduate advertising and public relations"	1
84	"post graduate degree"	1
85	"post graduate diploma rural management"	1
86	"prof ddr"	1
87	"professional certificate editing"	1
88	"professor chair"	1
89	"quantitative"	1
90	"social science"	1
91	"swiss certified tax expert"	1
92	"university saskatchewan"	1
93	"valuation"	1

6. Make decisions about tough cases

We assume that anyone listing professor has a PhD (i.e. "professor", "assistant professor", "prof", "associate professor"). We assume that anyone with a post graduate position has an MA.

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "prof", "phd"))
```

```
[1] "Observations changed: 10"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "post grad|postgrau", "ma"))
```

```
[1] "Observations changed: 4"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "graduate ", "ma"))
```

```
[1] "Observations changed: 6"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "frm", "ma"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "degree", "ba"))
```

```
[1] "Observations changed: 1"
```

```
  highest_degree <- highest_degree %>%
    mutate(degree_clean = simplify_response(degree_clean, "^diploma", "ba"))
```

```
[1] "Observations changed: 2"
```

Finally, we tidy up a few more edge cases.

```r
highest_degree <- highest_degree %>%
  mutate(
    degree_clean = case_match(
      degree_clean,
      "m s" ~ "ma",
      "acs" ~ "ba",
      "alm" ~ "ma",
      "m pub pol mgt" ~ "ma",
      "global studies ba" ~ "ba",
      "mla mms med" ~ "ma",
      .default = degree_clean
    )
  )
```

Here are the results before coarsening.

```r
highest_degree %>%
  count(degree_clean, sort = TRUE, wt = n) %>%
  print(n = Inf)
```

```
# A tibble: 66 x 2
   degree_clean                                                    n
   <chr>                                                       <dbl>
 1 "phd"                                                        3076
 2 "ma"                                                          520
 3 "ba"                                                          122
 4 ""                                                             14
 5 "jd"                                                           14
 6 "certificate"                                                   3
 7 "public policy"                                                 3
 8 "md"                                                            2
 9 "mpp"                                                           2
10 "agrégation"                                                    1
11 "agrégation des universités"                                    1
12 "alm finance"                                                   1
13 "and finance"                                                   1
14 "and finance aging"                                             1
15 "and japanese language literature"                              1
16 "applied nd"                                                    1
17 "applied ometrics"                                              1
18 "as business administration management"                         1
19 "assistant research fellow"                                     1
```

20	"b"	1
21	"be"	1
22	"business finance"	1
23	"business management"	1
24	"cert art intelligence business applications"	1
25	"certificate board leadership"	1
26	"certified business omist"	1
27	"cfa"	1
28	"cfa charterholder"	1
29	"computer science"	1
30	"continued education course"	1
31	"dimploma hotel and tourism management"	1
32	"education specialist special educ credential eds"	1
33	"educator certificate school guidance"	1
34	"environmental and policy"	1
35	"executive education evidence policy design"	1
36	"fellowship"	1
37	"finance"	1
38	"financial"	1
39	"fiscal crimes expert commercial money laundering and terrorism financ~	1
40	"foreign affairs"	1
41	"government and politics"	1
42	"graduated"	1
43	"international and finance"	1
44	"international applied specialist"	1
45	"llm"	1
46	"mbs"	1
47	"mdiv"	1
48	"me"	1
49	"med educational leadership policy and human development"	1
50	"methodology complex sales"	1
51	"mfa"	1
52	"mls"	1
53	"mmh"	1
54	"mom logistics and supply chain"	1
55	"mph"	1
56	"murp omic development"	1
57	"northwestern university"	1
58	"omic aanalysis specialist"	1
59	"omic sciences"	1
60	"organizational communication"	1
61	"p"	1
62	"quantitative"	1

```
63 "social science"                                                    1
64 "swiss certified tax expert"                                        1
65 "university saskatchewan"                                           1
66 "valuation"                                                         1
```

7. Coarsen to groups of interest

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "^mph|^mfa|^mdiv|^murp|^mmh|^mls|^
```

```
[1] "Observations changed: 13"
```

```
highest_degree <- highest_degree %>%
  mutate(degree_clean = simplify_response(degree_clean, "^be", "ma"))
```

```
[1] "Observations changed: 1"
```

Explore Results

At this point, most, if not all responses, don't answer the question about educational attainment. We will need to decide what to do about responses that don't answer the question of interest.

```
highest_degree %>%
  count(degree_clean, sort = TRUE, wt = n) %>%
  print(n = Inf)
```

```
# A tibble: 53 x 2
   degree_clean                                                        n
   <chr>                                                           <dbl>
 1 "phd"                                                            3076
 2 "ma"                                                              534
 3 "ba"                                                              122
 4 ""                                                                 14
 5 "jd"                                                               14
 6 "certificate"                                                       3
 7 "public policy"                                                     3
```

```
 8 "md"                                                                    2
 9 "agrégation"                                                            1
10 "agrégation des universités"                                            1
11 "alm finance"                                                           1
12 "and finance"                                                           1
13 "and finance aging"                                                     1
14 "and japanese language literature"                                      1
15 "applied nd"                                                            1
16 "applied ometrics"                                                      1
17 "as business administration management"                                 1
18 "assistant research fellow"                                             1
19 "b"                                                                     1
20 "business finance"                                                      1
21 "business management"                                                   1
22 "cert art intelligence business applications"                           1
23 "certificate board leadership"                                          1
24 "certified business omist"                                              1
25 "cfa"                                                                   1
26 "cfa charterholder"                                                     1
27 "computer science"                                                      1
28 "continued education course"                                            1
29 "dimploma hotel and tourism management"                                 1
30 "education specialist special educ credential eds"                      1
31 "educator certificate school guidance"                                  1
32 "environmental and policy"                                              1
33 "executive education evidence policy design"                            1
34 "fellowship"                                                            1
35 "finance"                                                               1
36 "financial"                                                             1
37 "fiscal crimes expert commercial money laundering and terrorism financ~ 1
38 "foreign affairs"                                                       1
39 "government and politics"                                               1
40 "graduated"                                                             1
41 "international and finance"                                             1
42 "international applied specialist"                                      1
43 "mom logistics and supply chain"                                        1
44 "northwestern university"                                               1
45 "omic aanalysis specialist"                                             1
46 "omic sciences"                                                         1
47 "organizational communication"                                          1
48 "p"                                                                     1
49 "quantitative"                                                          1
50 "social science"                                                        1
```

```
51 "swiss certified tax expert"                                  1
52 "university saskatchewan"                                     1
53 "valuation"                                                   1
```

```
highest_degree <- highest_degree %>%
  mutate(
    degree_simple = case_match(
      degree_clean,
      "phd" ~ "PhD or DBA",
      c("md", "jd") ~ "JD, MD, or other terminal degree",
      c("ma") ~ "Master's degree",
      c("ba") ~ "Bachelor's degree",
      .default = "other"
    )
  )
```

```
highest_degree %>%
  count(degree_simple, sort = TRUE, wt = n) %>%
  print(n = Inf)
```

```
# A tibble: 5 x 2
  degree_simple                       n
  <chr>                           <dbl>
1 PhD or DBA                       3076
2 Master's degree                   534
3 Bachelor's degree                 122
4 other                              65
5 JD, MD, or other terminal degree   16
```

Save Results

```
# detailed data
highest_degree %>%
  write_csv(here("data", "clean", "highest-degree_detailed.csv"))

# summarized data
highest_degree %>%
  count(degree_simple, sort = TRUE) %>%
  write_csv(here("data", "clean", "highest-degree_summarized.csv"))
```

Session Info

```
sessionInfo()
```

```
R version 4.2.2 (2022-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Monterey 12.2.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] readxl_1.4.2    here_1.0.1      lubridate_1.9.2 forcats_1.0.0
 [5] stringr_1.5.0   dplyr_1.1.1     purrr_1.0.1     readr_2.1.4
 [9] tidyr_1.3.0     tibble_3.2.1    ggplot2_3.4.2   tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] cellranger_1.1.0 pillar_1.9.0     compiler_4.2.2   tools_4.2.2
 [5] bit_4.0.5        digest_0.6.31    timechange_0.2.0 jsonlite_1.8.4
 [9] evaluate_0.20    lifecycle_1.0.3  gtable_0.3.3     pkgconfig_2.0.3
[13] rlang_1.1.0      cli_3.6.1        rstudioapi_0.14  parallel_4.2.2
[17] yaml_2.3.7       xfun_0.38        fastmap_1.1.1    withr_2.5.0
[21] knitr_1.42       generics_0.1.3   vctrs_0.6.2      hms_1.1.3
[25] bit64_4.0.5      rprojroot_2.0.3  grid_4.2.2       tidyselect_1.2.0
[29] glue_1.6.2       R6_2.5.1         fansi_1.0.4      vroom_1.6.1
[33] rmarkdown_2.21   tzdb_0.3.0       magrittr_2.0.3   scales_1.2.1
[37] htmltools_0.5.5  colorspace_2.1-0 utf8_1.2.3       stringi_1.7.12
[41] munsell_0.5.0    crayon_1.5.2
```