Outcome

QilinZhou

2023-12-12

```
outcome = read_dta("retention+difficulty.dta")
```

```
str(outcome)
```

```
## tibble [24,451,912 x 19] (S3: tbl_df/tbl/data.frame)
##  $ year    : num [1:24451912] 2007 2007 2007 2007 2007 ...
##   ..- attr(*, "label")= chr "survey year"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ serial  : num [1:24451912] 1 1 1 1 2 3 3 3 3 4 ...
##   ..- attr(*, "label")= chr "household serial number"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ month   : dbl+lbl [1:24451912] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
##    ..@ label       : chr "month"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:12] 1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:12] "january" "february" "march" "april" ...
##  $ hwtfinl : num [1:24451912] 3157 3157 3157 3157 3779 ...
##   ..- attr(*, "label")= chr "household weight, basic monthly"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ cpsid   : num [1:24451912] 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##   ..- attr(*, "label")= chr "cpsid, household record"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ asecflag: dbl+lbl [1:24451912] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label       : chr "flag for asec"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:2] 1 2
##    .. ..- attr(*, "names")= chr [1:2] "asec" "march basic"
##  $ hflag   : dbl+lbl [1:24451912] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label       : chr "flag for the 3/8 file 2014"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:2] 0 1
##    .. ..- attr(*, "names")= chr [1:2] "5/8 file" "3/8 file"
##  $ asecwth : num [1:24451912] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "annual social and economic supplement household weight"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ pernum  : num [1:24451912] 1 2 3 4 1 1 2 3 4 1 ...
##   ..- attr(*, "label")= chr "person number in sample unit"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ wtfinl  : num [1:24451912] 3285 3157 4598 2646 3779 ...
##   ..- attr(*, "label")= chr "final basic weight"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ cpsidv  : num [1:24451912] 2.01e+14 2.01e+14 2.01e+14 2.01e+14 2.01e+14 ...
##   ..- attr(*, "label")= chr "validated longitudinal identifier"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ cpsidp  : num [1:24451912] 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
##   ..- attr(*, "label")= chr "cpsid, person record"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ asecwt  : num [1:24451912] NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "annual social and economic supplement weight"
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ age     : dbl+lbl [1:24451912] 40, 39,  4,  7, 55, 50, 32, 10, 10, 60, 55, 56, 5...
##    ..@ label       : chr "age"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:100] 0 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:100] "under 1 year" "1" "2" "3" ...
##  $ sex     : dbl+lbl [1:24451912] 1, 2, 2, 1, 1, 2, 1, 2, 2, 2, 2, 1, 1, 2, 1, 2, 2...
##    ..@ label       : chr "sex"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:3] 1 2 9
##    .. ..- attr(*, "names")= chr [1:3] "male" "female" "niu"
##  $ race    : dbl+lbl [1:24451912] 200, 200, 200, 200, 100, 200, 200, 200, 200, 200,...
##    ..@ label       : chr "race"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:29] 100 200 300 650 651 652 700 801 802 803 ...
##    .. ..- attr(*, "names")= chr [1:29] "white" "black" "american indian/aleut/eskimo" "asian or paci
##  $ diffany : dbl+lbl [1:24451912] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label       : chr "any difficulty"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:3] 0 1 2
##    .. ..- attr(*, "names")= chr [1:3] "niu" "no difficulty" "has difficulty"
##  $ edgrade : dbl+lbl [1:24451912] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label       : chr "current level of school enrollment"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:25] 11 12 21 22 101 102 103 104 105 106 ...
##    .. ..- attr(*, "names")= chr [1:25] "nursery (pre-school, pre-k) part-day" "nursery (pre-school, 
##  $ edgrdly : dbl+lbl [1:24451912] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
##    ..@ label       : chr "level of school enrollment previous october"
##    ..@ format.stata: chr "%8.0g"
##    ..@ labels      : Named num [1:22] 10 20 101 102 103 104 105 106 107 108 ...
##    .. ..- attr(*, "names")= chr [1:22] "nursery (pre-school, pre-kindergarten)" "kindergarten" "1st 
```

```r
# Five-year-olds
filtered_data <- outcome %>%
  filter(age == 5)
```

```r
filtered_data <- filtered_data %>%
  filter(!is.na(edgrade), !is.na(edgrdly))
```

```r
# Level of school enrollment previous October
filtered_data <- filtered_data %>%
  mutate(category = case_when(
    as.character(edgrdly) == "10" ~ "Nursery (pre-school, pre-kindergarten)",
    as.character(edgrdly) == "20" ~ "Kindergarten",
    as.character(edgrdly) == "9998" ~ NA,
    as.character(edgrdly) == "9999" ~ "NIU",
    TRUE ~ "Others"  # Default case if none of the above
  ))
```

```r
filtered_data <- filtered_data %>%
  filter(category != "NIU")
```

```r
unique(filtered_data$category)
```

```
## [1] "Kindergarten"
## [2] "Nursery (pre-school, pre-kindergarten)"
## [3] "Others"
```

```r
# Current enrollment
filtered_data <- filtered_data %>%
  mutate(category_2 = case_when(
    as.character(edgrade) == "11" ~ "Nursery (pre-school, pre-K) part-day",
 as.character(edgrade)  == "12" ~ "Nursery (pre-school, pre-K) full-day",
 as.character(edgrade)  == "21" ~ "Kindergarten part-day",
 as.character(edgrade) == "22" ~ "Kindergarten full-day",
 as.character(edgrade) == "9998" ~ NA,
 as.character(edgrade) == "9999" ~ "NIU",
    TRUE ~ "Others"  # Default case if none of the above
  ))
```

```r
filtered_data <- filtered_data %>%
  filter(category_2 != "NIU")
```

```r
unique(filtered_data$category_2)
```

```
## [1] "Others"
## [2] "Kindergarten full-day"
## [3] "Kindergarten part-day"
## [4] "Nursery (pre-school, pre-K) part-day"
## [5] "Nursery (pre-school, pre-K) full-day"
```

```r
counts_prev_enroll <- filtered_data %>% count(category)
counts_prev_enroll
```

```
## # A tibble: 3 x 2
##   category                                   n
##   <chr>                                  <int>
## 1 Kindergarten                            1201
## 2 Nursery (pre-school, pre-kindergarten) 12410
## 3 Others                                   177
```

```r
counts_current_enroll <- filtered_data %>% count(category_2)
counts_current_enroll
```

```
## # A tibble: 5 x 2
##   category_2                               n
##   <chr>                                <int>
## 1 Kindergarten full-day                 8592
## 2 Kindergarten part-day                 2189
## 3 Nursery (pre-school, pre-K) full-day  1000
## 4 Nursery (pre-school, pre-K) part-day  1015
## 5 Others                                 992
```

```r
filtered_data <- filtered_data %>%
  mutate(retention_status = case_when(
    category == "Nursery (pre-school, pre-kindergarten)" &
    (category_2 == "Nursery (pre-school, pre-K) part-day" | category_2 == "Nursery (pre-school, pre-K) f
    !is.na(category) & !is.na(category_2) ~ "Not Retended",
    TRUE ~ NA
  ))
```

```
counts_retention <- filtered_data %>% count(retention_status)
counts_retention
```

```
## # A tibble: 2 x 2
##   retention_status     n
##   <chr>            <int>
## 1 Not Retended     11817
## 2 Retended          1971
```

```
filtered_data <- filtered_data %>%
  mutate(difficulty_status = case_when(
    as.character(diffany) == "1" ~ "No difficulty",
    as.character(diffany)  == "2" ~ "Has difficulty",
    as.character(diffany) == "0" ~ "NIU",
    TRUE ~ NA  # Default case if none of the above
  ))
```

```
counts_difficulty <- filtered_data %>% count(difficulty_status)
counts_difficulty
```

```
## # A tibble: 2 x 2
##   difficulty_status     n
##   <chr>             <int>
## 1 NIU               12739
## 2 <NA>               1049
```

```
# Kindergarten enrollment rate
# Preschool enrollment rate
# Retention among preschoolers

# Among all five-year-old children who have enrolled kindergarten
filtered_data <- filtered_data %>%
  mutate(enrolled_kindergarten = case_when(
    (category == "Kindergarten")|(category_2 == "Kindergarten full-day")|(category_2 == "Kindergarten pa
    TRUE ~ "No"
  ))
```

```
yearly_kindergarten_enrollment_rate <- filtered_data %>%
  group_by(year) %>%
  summarise(
    total_students = n(),
    kindergarten_enrolled = sum(enrolled_kindergarten == "Yes"),
    enrollment_rate = (kindergarten_enrolled / total_students) * 100
  )
```

```
print(yearly_kindergarten_enrollment_rate)
```

```
## # A tibble: 14 x 4
##     year total_students kindergarten_enrolled enrollment_rate
##    <dbl>          <int>                 <int>           <dbl>
##  1  2007           1049                   895            85.3
##  2  2008           1123                   975            86.8
##  3  2009           1119                   965            86.2
##  4  2010           1053                   889            84.4
##  5  2011           1105                   906            82.0
##  6  2012           1059                   900            85.0
##  7  2013           1005                   834            83.0
##  8  2014           1073                   914            85.2
##  9  2015            931                   770            82.7
## 10  2016            959                   791            82.5
## 11  2017            891                   734            82.4
## 12  2018            853                   720            84.4
## 13  2019            858                   714            83.2
## 14  2020            710                   597            84.1
```

```
# Among all five-year-old children who have enrolled preschool
filtered_data <- filtered_data %>%
  mutate(enrolled_preschool = case_when(
    (category == "Nursery (pre-school, pre-kindergarten)")|(category_2 == "Nursery (pre-school, pre-K) p
    TRUE ~ "No"
  ))
```

```
yearly_preschool_enrollment_rate <- filtered_data %>%
  group_by(year) %>%
  summarise(
    total_students = n(),
    preschool_enrolled = sum(enrolled_preschool == "Yes"),
    enrollment_rate = (preschool_enrolled / total_students) * 100
  )
```

```
print(yearly_preschool_enrollment_rate)
```

```
## # A tibble: 14 x 4
##     year total_students preschool_enrolled enrollment_rate
##    <dbl>          <int>              <int>           <dbl>
##  1  2007           1049                954            90.9
##  2  2008           1123                983            87.5
##  3  2009           1119               1016            90.8
##  4  2010           1053                933            88.6
##  5  2011           1105                995            90.0
##  6  2012           1059                955            90.2
##  7  2013           1005                901            89.7
##  8  2014           1073                962            89.7
##  9  2015            931                863            92.7
## 10  2016            959                886            92.4
## 11  2017            891                820            92.0
## 12  2018            853                782            91.7
## 13  2019            858                769            89.6
## 14  2020            710                635            89.4
```

```
# For preschoolers
yearly_retention_rate <- filtered_data %>%
  filter(enrolled_preschool == "Yes") %>%  # Filter for only preschoolers
  group_by(year) %>%
  summarise(
    total_preschoolers = n(),
    num_retention = sum(retention_status == "Retended"),
    retention_rate = (num_retention / total_preschoolers) * 100
  )
```

```
print(yearly_retention_rate)
```

```
## # A tibble: 14 x 4
##     year total_preschoolers num_retention retention_rate
##    <dbl>              <int>         <int>          <dbl>
##  1  2007                954           135           14.2
##  2  2008                983           126           12.8
##  3  2009               1016           130           12.8
##  4  2010                933           140           15.0
##  5  2011                995           186           18.7
##  6  2012                955           142           14.9
##  7  2013                901           161           17.9
##  8  2014                962           140           14.6
##  9  2015                863           149           17.3
## 10  2016                886           154           17.4
## 11  2017                820           145           17.7
## 12  2018                782           127           16.2
## 13  2019                769           137           17.8
## 14  2020                635            99           15.6
```

```
write.csv(filtered_data, "retention_diff.csv", row.names=FALSE)
```