02. Revise Data

Angel Huang, modified from Perry Haaland

11/29/2020

## Introduction

The purpose of the markdown document is to revise the existing data, to make it more suited for further analysis. In particular, one new variable is added (credits failed). All variables get R friendly names that will display better on graphics and generally require less typing.

This code produces a data set for common use. It is stored as an R objected in the data folder under the name revised\_df.RData.

## Load the previously prepared data

We have all working from the following common data set up to this point.

load("../../data/students\_df.RData")  
dim(students\_df)

## [1] 14238 37

We recall the variable names.

names(students\_df)

## [1] "UNIQUE\_ID" "STRM" "STRM\_DESC"   
## [4] "IS\_RETAINED" "IS\_RETAINED\_DESC" "RESIDENCY\_CODE"   
## [7] "RESIDENCY" "ACAD\_CAREER\_STUDENT" "CLASS\_LEVEL\_STUDENT"   
## [10] "IS\_DEGREE\_SEEKING" "INTENDED\_DECLARED" "INTENDED\_DECLARED\_DESC"  
## [13] "IS\_DBL\_MAJOR" "MAJ\_1\_SCHOOL\_CODE" "MAJ\_1\_SCHOOL\_NAME"   
## [16] "MAJ\_1\_DIV\_CODE" "MAJ\_1\_DIV\_NAME" "MAJ\_1\_PROGRAM\_DEGREE"   
## [19] "MAJ\_1\_PROGRAM\_CODE" "MAJ\_1\_PLAN\_CODE" "MAJ\_1\_PLAN\_DESC"   
## [22] "MAJ\_2\_SCHOOL\_CODE" "MAJ\_2\_SCHOOL\_NAME" "MAJ\_2\_DIV\_CODE"   
## [25] "MAJ\_2\_DIV\_NAME" "MAJ\_2\_PROGRAM\_DEGREE" "MAJ\_2\_PROGRAM\_CODE"   
## [28] "MAJ\_2\_PLAN\_CODE" "MAJ\_2\_PLAN\_DESC" "CUM\_RESIDENT\_TERMS\_BOT"  
## [31] "STU\_ORIG\_ENROLL\_STATUS" "FIRST\_STRM" "FIRST\_STRM\_DESC"   
## [34] "TRNSFR\_UNITS\_ENTRY" "TEST\_CREDIT\_ENTRY" "CREDIT\_HOURS\_ATT"   
## [37] "CREDIT\_HOURS\_EARNED"

### Variables for modeling

We are going to assign R friendly names, first.

* Retained = IS\_RETAINED: 0 for did not return, 1 for returned
* Residency = RESIDENCY: In-state, Out-of-state
* Class = CLASS\_LEVEL\_STUDENT: FR, SO, JR, and SR
* DoubleMajor = IS\_DBL\_MAJOR: 0 for No, 1 for Yes
* Enrollment = STU\_ORIG\_ENROLL\_STATUS: New Student or Transfer Student
* Credits = CREDIT\_HOURS\_EARNED
* CreditsFailed = CREDITS\_FAILED: CREDIT\_HOURS\_ATT - CREDIT\_HOURS\_EARNED
* MajorDivision = MAJ\_1\_DIV\_CODE, abbreviation for division of declared major
* Program = MAJ\_1\_PROGRAM\_DEGREE

As before, we start by making the class level an ordered factor based on student’s year in school.

class\_levels <- sort(unique(students\_df$CLASS\_LEVEL\_STUDENT))[c(1,3,2,4)]  
mod\_df <- students\_df  
mod\_df$CLASS\_LEVEL\_STUDENT <-  
 factor(students\_df$CLASS\_LEVEL\_STUDENT,   
 levels = class\_levels)

We also add CREDITS\_FAILED to the data. I like this variable a bit better than the credits earned because it seems to have a more easily grasped interpretation.

mod\_df <- mod\_df %>%  
 mutate(CREDITS\_FAILED = CREDIT\_HOURS\_ATT - CREDIT\_HOURS\_EARNED)

We also delete one very small group of students who don’t easily fit in with the original enrollment status.

mod\_df <- mod\_df %>%  
 filter(STU\_ORIG\_ENROLL\_STATUS != "Unclassified") %>%  
 mutate(STU\_ORIG\_ENROLL\_STATUS =   
 ifelse(STU\_ORIG\_ENROLL\_STATUS == "New Student",  
 "New Student",  
 "Transfer Student"))

Finally, for the purposes of this analysis and the ability to display the results, I rename many of the variables with R friendly and graphics suitable names.

revised\_df <- mod\_df %>%  
 select(Retained = IS\_RETAINED,  
 Class = CLASS\_LEVEL\_STUDENT,  
 Residency = RESIDENCY,  
 DoubleMajor = IS\_DBL\_MAJOR,  
 Enrollment = STU\_ORIG\_ENROLL\_STATUS,  
 Credits = CREDIT\_HOURS\_EARNED,  
 CreditsFailed = CREDITS\_FAILED,  
 MajorDiv = MAJ\_1\_DIV\_CODE,  
 Program = MAJ\_1\_PROGRAM\_DEGREE)

There are 14,171 students in our final data set.

nrow(revised\_df)

## [1] 14171

## Checking on the variables

revised\_df %>%  
 group\_by(Class) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 4 x 2  
## Class n  
## <fct> <int>  
## 1 FR 1842  
## 2 SO 4253  
## 3 JR 5248  
## 4 SR 2828

revised\_df %>%  
 group\_by(Residency) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 2 x 2  
## Residency n  
## <fct> <int>  
## 1 In-State 11705  
## 2 Out-of-State 2466

revised\_df %>%  
 group\_by(DoubleMajor) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 2 x 2  
## DoubleMajor n  
## <int> <int>  
## 1 0 10989  
## 2 1 3182

revised\_df %>%  
 group\_by(Enrollment) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 2 x 2  
## Enrollment n  
## <chr> <int>  
## 1 New Student 12723  
## 2 Transfer Student 1448

revised\_df %>%  
 group\_by(MajorDiv) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 15 x 2  
## MajorDiv n  
## <fct> <int>  
## 1 AH 58  
## 2 AS 839  
## 3 BU 1269  
## 4 DT 52  
## 5 ED 219  
## 6 FA 201  
## 7 HU 745  
## 8 IL 82  
## 9 MJ 1018  
## 10 NS 6567  
## 11 NU 423  
## 12 PH 312  
## 13 PR 74  
## 14 SB 2299  
## 15 UE 13

revised\_df %>%  
 group\_by(Program) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 12 x 2  
## Program n  
## <fct> <int>  
## 1 BA 5706  
## 2 BACH 839  
## 3 BAED 219  
## 4 BAMJ 676  
## 5 BFA 23  
## 6 BMUS 36  
## 7 BS 4512  
## 8 BSBA 1269  
## 9 BSIS 82  
## 10 BSN 423  
## 11 BSPH 312  
## 12 BSPHS 74

revised\_df %>%  
 group\_by(Credits) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 38 x 2  
## Credits n  
## <dbl> <int>  
## 1 0 614  
## 2 1 15  
## 3 2 4  
## 4 3 97  
## 5 4 24  
## 6 4.5 1  
## 7 5 7  
## 8 5.5 1  
## 9 6 139  
## 10 7 68  
## # ... with 28 more rows

revised\_df %>%  
 group\_by(CreditsFailed) %>%  
 summarize(n = n(), .groups = "drop")

## # A tibble: 28 x 2  
## CreditsFailed n  
## <dbl> <int>  
## 1 -2.5 1  
## 2 -2 10  
## 3 0 11651  
## 4 1 318  
## 5 1.5 77  
## 6 2 28  
## 7 3 933  
## 8 4 188  
## 9 4.5 3  
## 10 5 21  
## # ... with 18 more rows

## Saving the data

The revised file is saved in the data folder as revised\_df.RData.

save(revised\_df, file = "revised\_df.RData")