Capstone_wrangling_markdown

Grant Cox

November 3, 2018

Data Wrangling for Capstone: Overview

This project will potentially look at university football programs from a few angles. First, and more robustly, is through the revenue and subsidy information found in the 52 variable dataset which I'll refer to as the *Subsidized_programs Dataset*. Secondly, I'll look to win-loss records are reflected relative to some of the unique relationships found in the Subsidized_programs Dataset. Win-loss records are stored by year in 5 unique tables from the years 2010 through 2014. Year specific win-loss information was imported from 5 unique tabs in a csv file which will be referred to as the *WL Dataset* (or Win-Loss Dataset).

Removing unneccesary columns and combining datasets

My first task was to make the *Subsidized_programs Dataset* a bit less overwhelming by removing some of the variables I wouldn't need. Many columns were superfluous and many of the financial columns were repeated with similar information but accounting for inflation. While there may be a place for this, I started by eliminating columns I wouldn't need.

```r
library(tidyr)
library(dplyr)

df <- tbl_df(subsidized_programs)

df <- within(df, rm(X__1,city,state,url,nickname))
df <- df[, !grepl(pattern = "^inflation", colnames(df))] #remove inflation based columns

football_df <- df
write.csv(football_df, file = "football_df.csv")
```

From there, I needed to begin combining (in a couple ways). My *WL Dataset* existed as one workbook with 5 tabs, and my *WL Dataset* and my *Subsidized_programs Dataset* were separate entities that needed to be merged.

```r
wl10$year <- 2010
wl11$year <- 2011
wl12$year <- 2012
wl13$year <- 2013
wl14$year <- 2014

wl5year <- rbind(wl10, wl11, wl12, wl13, wl14)
View(wl5year[order(wl5year$Team, wl5year$year), ])

write.csv(wl5year, file = "wl5year.csv")
View(wl5year)
```

The rbind() command solved my first (admittedly easier) problem. I now had two table dataframes that I could merge, but with one big roadblock: university names were spelled out differently between these two. I knew that if I could align the two tables by university name, I could them merge them by name and by year, but this task proved more involved than anticipated.

Step one: add blank unitid column to wl5year.csv

```
wl5year$unitid <- NA

winlossid_tbl <- wl5year[,c(1,7)] %>% unique()
winlossid_tbl_alph <- winlossid_tbl[order(winlossid_tbl$Team), ]

View(winlossid_tbl_alph)
```

Step two: manually assign unitid

Easily the most arduous step. I created the above "winlossid_tbl_alph" to move through the win-loss schools one at a time, lining them up next to the schools and correlating unitids in the subsidized_program dataset.

```
wl5year$unitid[wl5year$Team == "Air Force"]   <- NA
wl5year$unitid[wl5year$Team == "Akron"]       <- 200800
wl5year$unitid[wl5year$Team == "Alabama"]     <- 100751
wl5year$unitid[wl5year$Team == "App State"]   <- 197869
wl5year$unitid[wl5year$Team == "Arizona"]     <- 104179
wl5year$unitid[wl5year$Team == "Arizona St"]  <- 104151
wl5year$unitid[wl5year$Team == "Arkansas"]    <- 106397
wl5year$unitid[wl5year$Team == "Akransas St"]   <- 106458
wl5year$unitid[wl5year$Team == "Army"]        <- NA
wl5year$unitid[wl5year$Team == "Auburn"]      <- 100858
wl5year$unitid[wl5year$Team == "Ball State"]  <- 150136
wl5year$unitid[wl5year$Team == "Baylor"]      <- NA
wl5year$unitid[wl5year$Team == "Boise State"] <- 142115
wl5year$unitid[wl5year$Team == "Boston Col"]  <- NA
wl5year$unitid[wl5year$Team == "Bowling Grn"] <- 201441
wl5year$unitid[wl5year$Team == "Buffalo"]     <- 196088
wl5year$unitid[wl5year$Team == "BYU"]         <- NA
wl5year$unitid[wl5year$Team == "California"]  <- 110635
wl5year$unitid[wl5year$Team == "Central FL"]  <- NA
wl5year$unitid[wl5year$Team == "Central Mich"]  <- 169248
wl5year$unitid[wl5year$Team == "Cincinnati"]  <- 201885
wl5year$unitid[wl5year$Team == "Clemson"]     <- 217882
wl5year$unitid[wl5year$Team == "Coastal Car"] <- NA
wl5year$unitid[wl5year$Team == "Colorado"]    <- 126614
wl5year$unitid[wl5year$Team == "Colorado St"] <- 126818
wl5year$unitid[wl5year$Team == "Connecticut"] <- 129020
wl5year$unitid[wl5year$Team == "Duke"]        <- NA
wl5year$unitid[wl5year$Team == "E Carolina"]  <- 198464
wl5year$unitid[wl5year$Team == "E Michigan"]  <- 169798
wl5year$unitid[wl5year$Team == "Fla Atlantic"]  <- 133669
wl5year$unitid[wl5year$Team == "Florida"]     <- 134130
wl5year$unitid[wl5year$Team == "Florida Intl"]  <- 133951
wl5year$unitid[wl5year$Team == "Florida St"]  <- 134097
wl5year$unitid[wl5year$Team == "Fresno St"]   <- 110556
wl5year$unitid[wl5year$Team == "GA Southern"] <- 139931
wl5year$unitid[wl5year$Team == "GA Tech"]     <- 139755
wl5year$unitid[wl5year$Team == "Georgia"]     <- 139959
wl5year$unitid[wl5year$Team == "Georgia State"] <- 139940
wl5year$unitid[wl5year$Team == "Hawaii"]      <- 141574
wl5year$unitid[wl5year$Team == "Houston"]     <- 225511
wl5year$unitid[wl5year$Team == "Illinois"]    <- 145637
wl5year$unitid[wl5year$Team == "Iowa State"]  <- 153603
```

```
wl5year$unitid[wl5year$Team == "Iowa"]        <- 153658
wl5year$unitid[wl5year$Team == "Kansas"]      <- 155317
wl5year$unitid[wl5year$Team == "Kansas St"]   <- 155399
wl5year$unitid[wl5year$Team == "Kent State"]  <- 203517
wl5year$unitid[wl5year$Team == "Kentucky"]    <- 157085
wl5year$unitid[wl5year$Team == "LA Lafayette"]  <- 160658
wl5year$unitid[wl5year$Team == "LA Monroe"]   <- NA
wl5year$unitid[wl5year$Team == "LA Tech"]     <- 159647
wl5year$unitid[wl5year$Team == "LSU"]          <- 159391
wl5year$unitid[wl5year$Team == "Liberty"]     <- NA
wl5year$unitid[wl5year$Team == "Louisville"]  <- 157289
wl5year$unitid[wl5year$Team == "Marshall"]    <- 237525
wl5year$unitid[wl5year$Team == "Maryland"]    <- 163286
wl5year$unitid[wl5year$Team == "Memphis"]     <- 220862
wl5year$unitid[wl5year$Team == "Miami (FL)"]  <- NA
wl5year$unitid[wl5year$Team == "Miami (OH)"]  <- NA
wl5year$unitid[wl5year$Team == "Michigan"]    <- 170976
wl5year$unitid[wl5year$Team == "Michigan St"] <- 171100
wl5year$unitid[wl5year$Team == "Middle Tenn"] <- 220978
wl5year$unitid[wl5year$Team == "Minnesota"]   <- 174066
wl5year$unitid[wl5year$Team == "Miss St"]     <- 176080
wl5year$unitid[wl5year$Team == "Mississippi"] <- 176017
wl5year$unitid[wl5year$Team == "Missouri"]    <- 178396
wl5year$unitid[wl5year$Team == "N Carolina"]  <- 199120
wl5year$unitid[wl5year$Team == "N Illinois"]  <- 147703
wl5year$unitid[wl5year$Team == "N Mex State"] <- 188030
wl5year$unitid[wl5year$Team == "Nebraska"]    <- 181464
wl5year$unitid[wl5year$Team == "Nevada"]      <- 182290
wl5year$unitid[wl5year$Team == "Navy"]        <- NA
wl5year$unitid[wl5year$Team == "NC State"]    <- 199193
wl5year$unitid[wl5year$Team == "New Mexico"]  <- 187985
wl5year$unitid[wl5year$Team == "North Texas"] <- 227216
wl5year$unitid[wl5year$Team == "Northwestern"]<- NA
wl5year$unitid[wl5year$Team == "Notre Dame"]  <- NA
wl5year$unitid[wl5year$Team == "Ohio"]        <- 204857
wl5year$unitid[wl5year$Team == "Ohio State"]  <- 204796
wl5year$unitid[wl5year$Team == "Oklahoma"]    <- 207500
wl5year$unitid[wl5year$Team == "Oklahoma St"] <- 207388
wl5year$unitid[wl5year$Team == "Old Dominion"] <- 232982
wl5year$unitid[wl5year$Team == "Oregon"]      <- 209551
wl5year$unitid[wl5year$Team == "Oregon St"]   <- 209542
wl5year$unitid[wl5year$Team == "Penn State"]  <- NA
wl5year$unitid[wl5year$Team == "Pittsburgh"]  <- NA
wl5year$unitid[wl5year$Team == "Purdue"]      <- 243780
wl5year$unitid[wl5year$Team == "Rice"]        <- NA
wl5year$unitid[wl5year$Team == "Rutgers"]     <- 186380
wl5year$unitid[wl5year$Team == "S Alabama"]   <- 102094
wl5year$unitid[wl5year$Team == "S Carolina"]  <- 218663
wl5year$unitid[wl5year$Team == "S Florida"]   <- 137351
wl5year$unitid[wl5year$Team == "S Methodist"] <- NA
wl5year$unitid[wl5year$Team == "S Mississippi"] <- 176372
wl5year$unitid[wl5year$Team == "San Diego St"]  <- NA
wl5year$unitid[wl5year$Team == "San Jose St"] <- 122755
```

```r
wl5year$unitid[wl5year$Team == "Stanford"]    <- NA
wl5year$unitid[wl5year$Team == "Syracuse"]    <- NA
wl5year$unitid[wl5year$Team == "Temple"]      <- NA
wl5year$unitid[wl5year$Team == "Tennessee"]   <- 221759
wl5year$unitid[wl5year$Team == "Texas"]       <- 228778
wl5year$unitid[wl5year$Team == "Texas A&M"]   <- 228723
wl5year$unitid[wl5year$Team == "Texas State"] <- 228459
wl5year$unitid[wl5year$Team == "Texas Tech"]  <- 229115
wl5year$unitid[wl5year$Team == "TX El Paso"]  <- 228796
wl5year$unitid[wl5year$Team == "Toledo"]      <- 206084
wl5year$unitid[wl5year$Team == "Troy"]        <- 102368
wl5year$unitid[wl5year$Team == "Tulane"]      <- NA
wl5year$unitid[wl5year$Team == "Tulsa"]       <- NA
wl5year$unitid[wl5year$Team == "U Mass"]      <- 166629
wl5year$unitid[wl5year$Team == "UAB"]         <- 100663
wl5year$unitid[wl5year$Team == "UCLA"]        <- 110662
wl5year$unitid[wl5year$Team == "UNLV"]        <- 182281
wl5year$unitid[wl5year$Team == "USC"]         <- NA
wl5year$unitid[wl5year$Team == "Utah"]        <- 230764
wl5year$unitid[wl5year$Team == "Utah State"]  <- 230728
wl5year$unitid[wl5year$Team == "VA Tech"]     <- 233921
wl5year$unitid[wl5year$Team == "Virginia"]    <- 234076
wl5year$unitid[wl5year$Team == "Vanderbilt"]  <- NA
wl5year$unitid[wl5year$Team == "W Kentucky"]  <- 157951
wl5year$unitid[wl5year$Team == "W Michigan"]  <- 172699
wl5year$unitid[wl5year$Team == "W Virginia"]  <- NA
wl5year$unitid[wl5year$Team == "Wake Forest"] <- NA
wl5year$unitid[wl5year$Team == "Wash State"]  <- 236939
wl5year$unitid[wl5year$Team == "Washingon"]   <- 236948
wl5year$unitid[wl5year$Team == "Wisconsin"]   <- 240444
wl5year$unitid[wl5year$Team == "Wyoming"]     <- 240727
```

Step three: prepare tables for merge and then merge

Needed to rename one column to account for repeated column names. Then merged the two tables via merge(), by the new “unitid” and by year.

```r
colnames(wl5year)[colnames(wl5year)=="Team"] <- "abbrev_name"

combined_df <- merge(wl5year, football_df, by = c("unitid", "year"), all = TRUE)

View(combined_df)
```