Capstone Project Data Wrangling

Varun Nadgir

August 8, 2017

The College Scorecard

Data Wrangling

The data available at The College Scorecard is composed of 19 .csv files - one for each academic year from 1996-'97 to 2014-'15. In these, there are about ~7,500 schools (rows) and 1,744 recorded data points (columns). To start, I wanted to merge all of these into one file and then I could break that file down into categorized, smaller datasets. Before merging, I would need to add a column to all 19 files that indicated the data collection year (taken from the filename). After that, I will be able to save the full dataset as a .csv locally for use in the future.

```r
# after importing the 19 .csv files, add the DATAYEAR column
MERGED1996_97_PP.csv <- MERGED1996_97_PP.csv %>% mutate(DATAYEAR = "1996-'97") # +18 more

# create and write 19 new .csv files which will be merged
write_csv(MERGED1996_97_PP.csv, path = "C:/Users/themi/Documents/Springboard/Foundations of Data Scienc

# path to new folder that holds multiple .csv files
data_folder <- "C:/Users/themi/Documents/Springboard/Foundations of Data Science/data/"

# create new list of all .csv files in folder
data_list <- list.files(path=data_folder, pattern="*.csv")

# read in each .csv file in file_list and rbind them into a data frame called fulldata
fulldata <-
  do.call("rbind",
          lapply(data_list,
                 function(x)
                 read.csv(paste(data_folder, x, sep=''),
                 stringsAsFactors = FALSE)))

# save fulldata as .csv locally
write_csv(fulldata, path = "C:/Users/themi/Documents/Springboard/Foundations of Data Science/data/fullda
```

Next, I want to break this large dataset down into smaller datasets, so that the data will be organized. At the moment, I have broken down the data into **Location** (city, zip, region), **Admission** (adm rate, demographics), **Financial** (cost, aid, debt), and **Education** (majors/minors, completion). I am in the process of studying the data dictionary, so that I may be able to rename columns as needed (since the original naming scheme is very technical and can be hard to read). Once my data is organized, I will be able to start plotting.