OfficalQCleaning

## Cleaning the data set

Needed Packages:

library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyverse)

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
v forcats 1.0.0 v readr 2.1.4  
v ggplot2 3.4.1 v stringr 1.5.0  
v lubridate 1.9.2 v tibble 3.2.1  
v purrr 1.0.1 v tidyr 1.3.0

-- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
x dplyr::filter() masks stats::filter()  
x dplyr::lag() masks stats::lag()  
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(lubridate)  
library(rio)  
library(writexl)  
setwd("C:/Users/rking/OneDrive/Documents/Econometrics/Data exploration project/Data\_Exploration\_Rawdata/")

Loading and joining trends data

file.list= list.files('Lab3\_Rawdata', pattern = "trends\_up\_to", full.names = "TRUE")  
dat <- import\_list(file.list, rbind = TRUE, fill = TRUE)

Changing the dates to aggregate by month

dat <- dat %>% mutate(monthorweek = str\_sub(monthorweek, 1, 10))  
  
  
dat <- dat%>% mutate(monthorweek = ymd(monthorweek))  
  
dat\_month <- dat %>% mutate(monthorweek = floor\_date(monthorweek, unit = "month"))

Standardizing the index’s by school and keyword to get them all on the same scale

standard\_df <- dat\_month %>%  
 group\_by(schname, keyword)%>%  
 mutate(stan\_index = (index- mean(index))/sd(index))  
  
#removing NA's from the data frame  
noNAs <- na.omit(standard\_df)

Aggregating standarized indexs to a school-month level

lld <- noNAs %>% group\_by(schname, monthorweek) %>%  
 summarise(new\_index = mean(stan\_index))

`summarise()` has grouped output by 'schname'. You can override using the  
`.groups` argument.

Importing the scorecard data and ID name link

scorecard <- import('Scorecard.csv')  
id\_link <- import('id\_name\_link.csv')  
#making scorecards column names lower cased  
names(scorecard) <- tolower(names(scorecard))  
  
#filtering for only bachleors degrees   
  
scorecard <- filter(scorecard, preddeg==3)

Getting rid of duplicate schools

id\_link <- id\_link %>% group\_by(schname) %>%  
 mutate(n=n())  
  
#filtering them out  
  
filtered\_id <- filter(id\_link, n==1)

Joining the three

id\_trends <- inner\_join(lld, filtered\_id)

Joining with `by = join\_by(schname)`

#now joining those to the scorecard  
  
final\_piece <- inner\_join(id\_trends, scorecard, by = 'unitid')

Writing my clean df to an excel doc to use for analysis

write\_xlsx(final\_piece, "final\_data.xlsx")