TCC IDHM Engenharia de dados

Matheus Assis e Oliveira

2023-10-10

carregar pacotes para a Engenharia de dados.

seguindo começamos a importar os dados em arquivo excel: xlsx, e converter anos a coluna, limpar o X a frente dos anos e converter anos para numérico.

renda\_per\_capita <- read\_excel("renda.per.capita.xlsx")  
#usar gather para converter anos à coluna  
renda\_per\_capita <- gather(renda\_per\_capita, ano, renda\_per\_capita, X2012:X2021, convert = TRUE)  
#Limpar o X a frente do ano  
renda\_per\_capita$ano <- gsub('X', '', renda\_per\_capita$ano)  
#converter ano de character para numeric  
renda\_per\_capita <- transform(renda\_per\_capita, ano = as.numeric(ano))  
renda\_per\_capita <- transform(renda\_per\_capita, renda\_per\_capita = as.numeric(renda\_per\_capita))

em seguida repete para demais arquivos com as seguintes variáveis.

sub\_esco\_pop <- read\_excel("sub.esco.pop.xlsx")  
sub\_esco\_pop <- gather(sub\_esco\_pop, ano, sub\_esco\_pop, X2012:X2021, convert = TRUE)  
sub\_esco\_pop$ano <- gsub('X', '', sub\_esco\_pop$ano)  
sub\_esco\_pop <- transform(sub\_esco\_pop, ano = as.numeric(ano))  
sub\_esco\_pop <- transform(sub\_esco\_pop, sub\_esco\_pop = as.numeric(sub\_esco\_pop))  
  
  
  
sub\_freq\_esco <- read\_excel("sub.freq.esco.xlsx")  
sub\_freq\_esco <- gather(sub\_freq\_esco, ano, sub\_freq\_esco, X2012:X2021, convert = TRUE)  
sub\_freq\_esco$ano <- gsub('X', '', sub\_freq\_esco$ano)  
sub\_freq\_esco <- transform(sub\_freq\_esco, ano = as.numeric(ano))  
sub\_freq\_esco <- transform(sub\_freq\_esco, sub\_freq\_esco = as.numeric(sub\_freq\_esco))  
  
  
  
  
esperança\_de\_vida <- read\_excel("esperança.de.vida.xlsx")  
esperança\_de\_vida <- gather(esperança\_de\_vida, ano, esperança\_de\_vida, X2012:X2021, convert = TRUE)  
esperança\_de\_vida$ano <- gsub('X', '', esperança\_de\_vida$ano)  
esperança\_de\_vida <- transform(esperança\_de\_vida, ano = as.numeric(ano))  
esperança\_de\_vida <- transform(esperança\_de\_vida, esperança\_de\_vida = as.numeric(esperança\_de\_vida))  
  
  
  
porcent\_pobres <- read\_excel("porcent\_pobres.xlsx")  
porcent\_pobres <- gather(porcent\_pobres, ano, porcent\_pobres, X2012:X2021, convert = TRUE)  
porcent\_pobres$ano <- gsub('X', '', porcent\_pobres$ano)  
porcent\_pobres <- transform(porcent\_pobres, ano = as.numeric(ano))  
porcent\_pobres <- transform(porcent\_pobres, porcent\_pobres = as.numeric(porcent\_pobres))  
  
  
  
população\_total <- read\_excel("população\_total.xlsx")  
população\_total <- gather(população\_total, ano, população\_total, X2012:X2021, convert = TRUE)  
população\_total$ano <- gsub('X', '', população\_total$ano)  
população\_total <- transform(população\_total, ano = as.numeric(ano))  
população\_total <- transform(população\_total, população\_total = as.numeric(população\_total))  
  
  
  
mortalidade\_infantil <- read\_excel("mortalidade\_infantil.xlsx")  
mortalidade\_infantil <- gather(mortalidade\_infantil, ano, mortalidade\_infantil, X2012:X2021, convert = TRUE)  
mortalidade\_infantil$ano <- gsub('X', '', mortalidade\_infantil$ano)  
mortalidade\_infantil <- transform(mortalidade\_infantil, ano = as.numeric(ano))  
mortalidade\_infantil <- transform(mortalidade\_infantil, mortalidade\_infantil = as.numeric(mortalidade\_infantil))  
  
  
  
media\_anos\_de\_estudo <- read\_excel("media\_anos\_de\_estudo.xlsx")  
media\_anos\_de\_estudo <- gather(media\_anos\_de\_estudo, ano, media\_anos\_de\_estudo, X2012:X2021, convert = TRUE)  
media\_anos\_de\_estudo$ano <- gsub('X', '', media\_anos\_de\_estudo$ano)  
media\_anos\_de\_estudo <- transform(media\_anos\_de\_estudo, ano = as.numeric(ano))  
media\_anos\_de\_estudo <- transform(media\_anos\_de\_estudo, media\_anos\_de\_estudo = as.numeric(media\_anos\_de\_estudo))  
  
  
  
indice\_gini <- read\_excel("indice\_gini.xlsx")  
indice\_gini <- gather(indice\_gini, ano, indice\_gini, X2012:X2021, convert = TRUE)  
indice\_gini$ano <- gsub('X', '', indice\_gini$ano)  
indice\_gini <- transform(indice\_gini, ano = as.numeric(ano))  
indice\_gini <- transform(indice\_gini, indice\_gini = as.numeric(indice\_gini))  
  
  
  
ind\_theil\_L <- read\_excel("ind\_theil\_L.xlsx")  
ind\_theil\_L <- gather(ind\_theil\_L, ano, ind\_theil\_L, X2012:X2021, convert = TRUE)  
ind\_theil\_L$ano <- gsub('X', '', ind\_theil\_L$ano)  
ind\_theil\_L <- transform(ind\_theil\_L, ano = as.numeric(ano))  
ind\_theil\_L <- transform(ind\_theil\_L, ind\_theil\_L = as.numeric(ind\_theil\_L))  
  
  
  
analfabetismo\_25\_anos\_mais\_idade <- read\_excel("analfabetismo\_25\_anos\_mais\_idade.xlsx")  
analfabetismo\_25\_anos\_mais\_idade <- gather(analfabetismo\_25\_anos\_mais\_idade, ano, analfabetismo\_25\_anos\_mais\_idade, X2012:X2021, convert = TRUE)  
analfabetismo\_25\_anos\_mais\_idade$ano <- gsub('X', '', analfabetismo\_25\_anos\_mais\_idade$ano)  
analfabetismo\_25\_anos\_mais\_idade <- transform(analfabetismo\_25\_anos\_mais\_idade, ano = as.numeric(ano))  
analfabetismo\_25\_anos\_mais\_idade <- transform(analfabetismo\_25\_anos\_mais\_idade, analfabetismo\_25\_anos\_mais\_idade = as.numeric(analfabetismo\_25\_anos\_mais\_idade))  
  
  
  
analfabetismo\_18\_anos\_mais\_idade <- read\_excel("analfabetismo\_18\_anos\_mais\_idade.xlsx")  
analfabetismo\_18\_anos\_mais\_idade <- gather(analfabetismo\_18\_anos\_mais\_idade, ano, analfabetismo\_18\_anos\_mais\_idade, X2012:X2021, convert = TRUE)  
analfabetismo\_18\_anos\_mais\_idade$ano <- gsub('X', '', analfabetismo\_18\_anos\_mais\_idade$ano)  
analfabetismo\_18\_anos\_mais\_idade <- transform(analfabetismo\_18\_anos\_mais\_idade, ano = as.numeric(ano))  
analfabetismo\_18\_anos\_mais\_idade <- transform(analfabetismo\_18\_anos\_mais\_idade, analfabetismo\_18\_anos\_mais\_idade = as.numeric(analfabetismo\_18\_anos\_mais\_idade))  
  
  
  
analfabetismo\_15\_anos\_mais\_idade <- read\_excel("analfabetismo\_15\_anos\_mais\_idade.xlsx")  
analfabetismo\_15\_anos\_mais\_idade <- gather(analfabetismo\_15\_anos\_mais\_idade, ano, analfabetismo\_15\_anos\_mais\_idade, X2012:X2021, convert = TRUE)  
analfabetismo\_15\_anos\_mais\_idade$ano <- gsub('X', '', analfabetismo\_15\_anos\_mais\_idade$ano)  
analfabetismo\_15\_anos\_mais\_idade <- transform(analfabetismo\_15\_anos\_mais\_idade, ano = as.numeric(ano))  
analfabetismo\_15\_anos\_mais\_idade <- transform(analfabetismo\_15\_anos\_mais\_idade, analfabetismo\_15\_anos\_mais\_idade = as.numeric(analfabetismo\_15\_anos\_mais\_idade))  
  
  
  
IDHM <- read\_excel("IDHM.xlsx")  
IDHM <- gather(IDHM, ano, IDHM, X2012:X2021, convert = TRUE)  
IDHM$ano <- gsub('X', '', IDHM$ano)  
IDHM <- transform(IDHM, ano = as.numeric(ano))  
IDHM <- transform(IDHM, IDHM = as.numeric(IDHM))

para cada indicator, conta número de linhas e o número total de NULLS e divide NULLS pela linha para obter a % de NULLS para indicador

print(paste0("renda\_per\_capita"))

## [1] "renda\_per\_capita"

renda\_per\_capita.na <- as.data.frame(sum(is.na(renda\_per\_capita$renda\_per\_capita)))  
renda\_per\_capita.n <- as.data.frame(nrow(renda\_per\_capita))  
renda\_per\_capita.na$`sum(is.na(renda\_per\_capita$renda\_per\_capita))`/renda\_per\_capita.n$`nrow(renda\_per\_capita)`\*100

## [1] 0

print(paste0("sub\_esco\_pop"))

## [1] "sub\_esco\_pop"

sub\_esco\_pop.na <- as.data.frame(sum(is.na(sub\_esco\_pop$sub\_esco\_pop)))  
sub\_esco\_pop.n <- as.data.frame(nrow(sub\_esco\_pop))  
sub\_esco\_pop.na$`sum(is.na(sub\_esco\_pop$sub\_esco\_pop))`/sub\_esco\_pop.n$`nrow(sub\_esco\_pop)`\*100

## [1] 0

print(paste0("sub\_freq\_esco"))

## [1] "sub\_freq\_esco"

sub\_freq\_esco.na <- as.data.frame(sum(is.na(sub\_freq\_esco$sub\_freq\_esco)))  
sub\_freq\_esco.n <- as.data.frame(nrow(sub\_freq\_esco))  
sub\_freq\_esco.na$`sum(is.na(sub\_freq\_esco$sub\_freq\_esco))`/sub\_freq\_esco.n$`nrow(sub\_freq\_esco)`\*100

## [1] 0

print(paste0("esperança\_de\_vida"))

## [1] "esperança\_de\_vida"

esperança\_de\_vida.na <- as.data.frame(sum(is.na(esperança\_de\_vida$esperança\_de\_vida)))  
esperança\_de\_vida.n <- as.data.frame(nrow(esperança\_de\_vida))  
esperança\_de\_vida.na$`sum(is.na(esperança\_de\_vida$esperança\_de\_vida))`/esperança\_de\_vida.n$`nrow(esperança\_de\_vida)`\*100

## [1] 0

print(paste0("porcent\_pobres"))

## [1] "porcent\_pobres"

porcent\_pobres.na <- as.data.frame(sum(is.na(porcent\_pobres$porcent\_pobres)))  
porcent\_pobres.n <- as.data.frame(nrow(porcent\_pobres))  
porcent\_pobres.na$`sum(is.na(rporcent\_pobres$porcent\_pobres))`/porcent\_pobres.n$`nrow(porcent\_pobres)`\*100

## numeric(0)

print(paste0("população\_total"))

## [1] "população\_total"

população\_total.na <- as.data.frame(sum(is.na(população\_total$população\_total)))  
população\_total.n <- as.data.frame(nrow(população\_total))  
população\_total.na$`sum(is.na(população\_total$população\_total))`/população\_total.n$`nrow(população\_total)`\*100

## [1] 0

print(paste0("mortalidade\_infantil"))

## [1] "mortalidade\_infantil"

mortalidade\_infantil.na <- as.data.frame(sum(is.na(mortalidade\_infantil$mortalidade\_infantil)))  
mortalidade\_infantil.n <- as.data.frame(nrow(mortalidade\_infantil))  
mortalidade\_infantil.na$`sum(is.na(mortalidade\_infantil$mortalidade\_infantil))`/mortalidade\_infantil.n$`nrow(mortalidade\_infantil)`\*100

## [1] 0

print(paste0("media\_anos\_de\_estudo"))

## [1] "media\_anos\_de\_estudo"

media\_anos\_de\_estudo.na <- as.data.frame(sum(is.na(media\_anos\_de\_estudo$media\_anos\_de\_estudo)))  
media\_anos\_de\_estudo.n <- as.data.frame(nrow(media\_anos\_de\_estudo))  
media\_anos\_de\_estudo.na$`sum(is.na(media\_anos\_de\_estudo$media\_anos\_de\_estudo))`/media\_anos\_de\_estudo.n$`nrow(media\_anos\_de\_estudo)`\*100

## [1] 0

print(paste0("indice\_gini"))

## [1] "indice\_gini"

indice\_gini.na <- as.data.frame(sum(is.na(indice\_gini$indice\_gini)))  
indice\_gini.n <- as.data.frame(nrow(indice\_gini))  
indice\_gini.na$`sum(is.na(indice\_gini$indice\_gini))`/indice\_gini.n$`nrow(indice\_gini)`\*100

## [1] 0

print(paste0("ind\_theil\_L"))

## [1] "ind\_theil\_L"

ind\_theil\_L.na <- as.data.frame(sum(is.na(ind\_theil\_L$ind\_theil\_L)))  
ind\_theil\_L.n <- as.data.frame(nrow(ind\_theil\_L))  
ind\_theil\_L.na$`sum(is.na(ind\_theil\_L$ind\_theil\_L))`/ind\_theil\_L.n$`nrow(ind\_theil\_L)`\*100

## [1] 0

print(paste0("analfabetismo\_25\_anos\_mais\_idade"))

## [1] "analfabetismo\_25\_anos\_mais\_idade"

analfabetismo\_25\_anos\_mais\_idade.na <- as.data.frame(sum(is.na(analfabetismo\_25\_anos\_mais\_idade$analfabetismo\_25\_anos\_mais\_idade)))  
analfabetismo\_25\_anos\_mais\_idade.n <- as.data.frame(nrow(analfabetismo\_25\_anos\_mais\_idade))  
analfabetismo\_25\_anos\_mais\_idade.na$`sum(is.na(analfabetismo\_25\_anos\_mais\_idade$analfabetismo\_25\_anos\_mais\_idade))`/analfabetismo\_25\_anos\_mais\_idade.n$`nrow(analfabetismo\_25\_anos\_mais\_idade)`\*100

## [1] 0

print(paste0("analfabetismo\_18\_anos\_mais\_idade"))

## [1] "analfabetismo\_18\_anos\_mais\_idade"

analfabetismo\_18\_anos\_mais\_idade.na <- as.data.frame(sum(is.na(analfabetismo\_18\_anos\_mais\_idade$analfabetismo\_18\_anos\_mais\_idade)))  
analfabetismo\_18\_anos\_mais\_idade.n <- as.data.frame(nrow(analfabetismo\_18\_anos\_mais\_idade))  
analfabetismo\_18\_anos\_mais\_idade.na$`sum(is.na(analfabetismo\_18\_anos\_mais\_idade$analfabetismo\_18\_anos\_mais\_idade))`/analfabetismo\_18\_anos\_mais\_idade.n$`nrow(analfabetismo\_18\_anos\_mais\_idade)`\*100

## [1] 0

print(paste0("analfabetismo\_15\_anos\_mais\_idade"))

## [1] "analfabetismo\_15\_anos\_mais\_idade"

analfabetismo\_15\_anos\_mais\_idade.na <- as.data.frame(sum(is.na(analfabetismo\_15\_anos\_mais\_idade$analfabetismo\_15\_anos\_mais\_idade)))  
analfabetismo\_15\_anos\_mais\_idade.n <- as.data.frame(nrow(analfabetismo\_15\_anos\_mais\_idade))  
analfabetismo\_15\_anos\_mais\_idade.na$`sum(is.na(analfabetismo\_15\_anos\_mais\_idade$analfabetismo\_15\_anos\_mais\_idade))`/analfabetismo\_15\_anos\_mais\_idade.n$`nrow(analfabetismo\_15\_anos\_mais\_idade)`\*100

## [1] 0

print(paste0("IDHM"))

## [1] "IDHM"

IDHM.na <- as.data.frame(sum(is.na(IDHM$IDHM)))  
IDHM.n <- as.data.frame(nrow(IDHM))  
IDHM.na$`sum(is.na(IDHM$IDHM))`/IDHM.n$`nrow(IDHM)`\*100

## [1] 0

#criar data frame único com os indicadores IDHM.key e IDHM.ind

IDHM.key = renda\_per\_capita  
IDHM.key <- join(IDHM.key, sub\_esco\_pop, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, sub\_freq\_esco, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, esperança\_de\_vida, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, porcent\_pobres, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, população\_total, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, mortalidade\_infantil, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, media\_anos\_de\_estudo, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, indice\_gini, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, ind\_theil\_L, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, analfabetismo\_25\_anos\_mais\_idade, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, analfabetismo\_18\_anos\_mais\_idade, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, analfabetismo\_15\_anos\_mais\_idade, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))  
IDHM.key = join(IDHM.key, IDHM, by = c("ano" = "ano", "Territorialidades" = "Territorialidades"))

key.ind = IDHM.key  
  
  
  
# Verificar que o número de estados continua os mesmos e o data frame está correto  
  
  
sapply(key.ind, function(x) length(unique(x)))

## Territorialidades ano   
## 28 10   
## renda\_per\_capita sub\_esco\_pop   
## 280 179   
## sub\_freq\_esco esperança\_de\_vida   
## 156 238   
## porcent\_pobres população\_total   
## 261 280   
## mortalidade\_infantil media\_anos\_de\_estudo   
## 255 202   
## indice\_gini ind\_theil\_L   
## 134 182   
## analfabetismo\_25\_anos\_mais\_idade analfabetismo\_18\_anos\_mais\_idade   
## 264 252   
## analfabetismo\_15\_anos\_mais\_idade IDHM   
## 248 135

str(key.ind)

## 'data.frame': 280 obs. of 16 variables:  
## $ Territorialidades : chr "Brasil" "Acre" "Alagoas" "Amapá" ...  
## $ ano : num 2012 2012 2012 2012 2012 ...  
## $ renda\_per\_capita : num 759 517 395 528 559 ...  
## $ sub\_esco\_pop : num 0.606 0.59 0.487 0.67 0.613 0.51 0.54 0.765 0.613 0.619 ...  
## $ sub\_freq\_esco : num 0.731 0.681 0.645 0.653 0.642 0.639 0.742 0.77 0.735 0.741 ...  
## $ esperança\_de\_vida : num 74.5 72.5 70 72.8 70.8 ...  
## $ porcent\_pobres : num 11.4 23.8 23.4 18.4 22.2 ...  
## $ população\_total : num 1.98e+08 7.77e+05 3.22e+06 7.21e+05 3.54e+06 ...  
## $ mortalidade\_infantil : num 15.8 20.2 26.1 24.3 20.9 ...  
## $ media\_anos\_de\_estudo : num 8.56 7.72 6.8 9.09 8.63 ...  
## $ indice\_gini : num 0.54 0.566 0.503 0.528 0.589 0.563 0.545 0.601 0.489 0.474 ...  
## $ ind\_theil\_L : num 0.526 0.585 0.447 0.483 0.619 0.571 0.54 0.664 0.411 0.383 ...  
## $ analfabetismo\_25\_anos\_mais\_idade: num 10.22 18.22 24.22 7.93 9.46 ...  
## $ analfabetismo\_18\_anos\_mais\_idade: num 8.75 14.72 20.54 6.37 7.89 ...  
## $ analfabetismo\_15\_anos\_mais\_idade: num 8.21 13.48 18.97 5.76 7.22 ...  
## $ IDHM : num 0.746 0.701 0.651 0.707 0.691 0.678 0.701 0.825 0.758 0.744 ...

#Key.ind é o data frame usado na etapa de análise exploratória de dados. # This final data frame (key.ind) will be used for the next step, exploratory data analysis.