# Regrouper par variable les donnees salariales des femmes et des hommes.

# Une base correspond a une categorie de travail d'homme / femme avec les annees comme variable

# On peut calculer directement le ratio salaire homme divise par salaire femme

# Importation des packages

library(readxl)

library(dplyr)

# -------------------------------------------------- #

# Importation des données

# -------------------------------------------------- #

setwd("D:/BUT SD - Année 2/Programmation R/Projet/inegaliteSalariale")

# 2012 - 2015

for(annee in c(2012:2015)){

assign(paste("df",annee, sep = ""),read\_xls(path = paste("data/",annee,".xls", sep = ""),

sheet = "DEP",

skip = 5))

}

# 2016 - 2021

for(annee in c(2016:2021)){

assign(paste("df",annee, sep = ""),read\_xlsx(path = paste("data/",annee,".xlsx", sep = ""),

sheet = "DEP",

skip = 5))

}

# -------------------------------------------------- #

# Sélection des variables qui nous intéressent

# -------------------------------------------------- #

variables = c("SNHMF","SNHMH","SNHMFC","SNHMHC","SNHMFP","SNHMHP","SNHMFE","SNHMHE","SNHMFO","SNHMHO")

cles = c("CODGEO","LIBGEO")

for(annee in c(12:21)){

assign(paste("df20", annee, sep = ""), get(paste("df20", annee, sep = "")) %>%

select(paste(variables, annee, sep = ""),

cles))

}

# -------------------------------------------------- #

# Création des bases pour les prédictions

# -------------------------------------------------- #

colonnes = c("CODGEO", "LIBGEO", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021")

for(var in variables){

assign(var, as.data.frame(df2012[c("CODGEO","LIBGEO")]))

for(annee in c(12:21)) {

suppressMessages({

assign(var, left\_join(get(var),

get(paste("df20", annee, sep = ""))[c("CODGEO","LIBGEO",paste(var, annee, sep = ""))]))

})

}

assign(var, setNames(get(var), colonnes))

}

# -------------------------------------------------- #

# Netoyage des variables

# -------------------------------------------------- #

variables\_supr = setdiff(ls(), variables)

rm(list = variables\_supr, variables\_supr)

# -------------------------------------------------- #

# Les prédictions

# -------------------------------------------------- #

bases = c("SNHMF","SNHMH","SNHMFC","SNHMHC","SNHMFP","SNHMHP","SNHMFE","SNHMHE","SNHMFO","SNHMHO")

cles = c("CODGEO","LIBGEO")

# Calcul des prédictions au delà de 2021 à partir de l'écart moyen entre les années connues

Moyenne\_ecart = function(base){

base["Moyenne\_ecart"] = (base["2013"] - base["2012"]+

base["2014"] - base["2013"]+

base["2015"] - base["2014"]+

base["2016"] - base["2015"]+

base["2017"] - base["2016"]+

base["2018"] - base["2017"]+

base["2019"] - base["2018"]+

base["2020"] - base["2019"]+

base["2021"] - base["2020"])/9

return(base)

}

for(base in bases){

# Il y a quelques valeurs manquantes présente en 2016

# On les estime à l'aide d'un modèle polynomial

donnees\_completes = get(base)[!is.na(get(base)["2016"]),]

donnees\_manquantes = get(base)[is.na(get(base)["2016"]),]

# Création du modèle à partir des individus sans valeur manquante

modele <- lm(`2016` ~ poly(`2012`, `2013`, `2014`, `2015`, `2017`, `2018`, `2019`, `2020`, `2021`,

degree = 2),

data = donnees\_completes)

# Prédiction à partir du modèle polynomial

predictions = predict(modele, newdata = donnees\_manquantes)

donnees\_manquantes["2016"] = predictions

# Ajout des prédictions de 2016 dans une nouvelle base

donnees\_finales = rbind(donnees\_completes, donnees\_manquantes)

esti = paste(base, "\_ESTI", sep = "")

assign(esti, get(base)[cles])

assign(esti, left\_join(get(esti),

donnees\_finales))

}

SNHMF\_ESTI <- SNHMF\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMFC\_ESTI <- SNHMFC\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMFE\_ESTI <- SNHMFE\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMFO\_ESTI <- SNHMFO\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMFP\_ESTI <- SNHMFP\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMH\_ESTI <- SNHMH\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMHC\_ESTI <- SNHMHC\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMHE\_ESTI <- SNHMHE\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMHO\_ESTI <- SNHMHO\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

SNHMHP\_ESTI <- SNHMHP\_ESTI %>% mutate(`2022` = NA, `2023` = NA, `2024` = NA, `2025` = NA, `2026` = NA, `2027` = NA, `2028` = NA, `2029` = NA, `2030` = NA,`2031` = NA, `2032` = NA, `2033` = NA, `2034` = NA, `2035` = NA, `2036` = NA, `2037` = NA, `2038` = NA, `2039` = NA,`2040` = NA)

for(base in bases){

# Calcule de l'écart moyen entre les années par département

esti = paste(base, "\_ESTI", sep = "")

assign(esti, Moyenne\_ecart(get(esti)))

}

for(predic in c(1:19)){SNHMF\_ESTI[as.character(2021+predic)] = SNHMF\_ESTI["2021"] + SNHMF\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMFC\_ESTI[as.character(2021+predic)] = SNHMFC\_ESTI["2021"] + SNHMFC\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMFE\_ESTI[as.character(2021+predic)] = SNHMFE\_ESTI["2021"] + SNHMFE\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMFO\_ESTI[as.character(2021+predic)] = SNHMFO\_ESTI["2021"] + SNHMFO\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMFP\_ESTI[as.character(2021+predic)] = SNHMFP\_ESTI["2021"] + SNHMFP\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMH\_ESTI[as.character(2021+predic)] = SNHMH\_ESTI["2021"] + SNHMH\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMHC\_ESTI[as.character(2021+predic)] = SNHMHC\_ESTI["2021"] + SNHMHC\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMHE\_ESTI[as.character(2021+predic)] = SNHMHE\_ESTI["2021"] + SNHMHE\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMHO\_ESTI[as.character(2021+predic)] = SNHMHO\_ESTI["2021"] + SNHMHO\_ESTI["Moyenne\_ecart"] \* predic}

for(predic in c(1:19)){SNHMHP\_ESTI[as.character(2021+predic)] = SNHMHP\_ESTI["2021"] + SNHMHP\_ESTI["Moyenne\_ecart"] \* predic}

# -------------------------------------------------- #

# Création des tables finales

# -------------------------------------------------- #

for(annee in c(2012:2040)){

base = paste("Salaire", annee, sep = "")

assign(base, SNHMF[cles])

compteur = 1

for(var in bases){

suppressMessages({

assign(base, left\_join(get(base),

get(paste(var, "\_ESTI", sep = ""))[c(cles,as.character(annee))]))

assign(base, setNames(get(base), c(cles, bases[1:compteur])))

compteur = compteur + 1

})

}

}

# Ajout de la longitude et latitude et calcul ratio

coords = read.csv("geo.csv",

sep = "\t")

colnames(coords)[4] = "CODGEO"

for(annee in c(2012:2040)){

base = get(paste("Salaire", annee, sep = ""))

for(var in c(1:5)){

base[paste(bases[var\*2], "r",sep = "")] = 1 - base[bases[var\*2 - 1]] / base[bases[var\*2]]

}

assign(paste("Salaire", annee, sep = ""), left\_join(get(paste("Salaire", annee, sep = "")), coords))

}

# -------------------------------------------------- #

# Sauvegarde des bases de données

# -------------------------------------------------- #

for(annee in c(2012:2040)){

base = paste("Salaire", annee, sep = "")

test = get(base)

save(test, file = paste("sorties/", annee, ".RData", sep = ""))

}

Salaire2030["SNHMr"] = Salaire2030$SNHMF / Salaire2030$SNHMH - 1

write.csv(Salaire2030, file = "Sorties/2030.csv", row.names = FALSE)