R Notebook for creation Grampian Data set

Rscript created by  
Author: Bernhard Scheliga  
Date: 2020-06-11

Data set created by  
Author: Bernhard Scheliga

## [1] "Dataset version: 0.1"

## [1] "Date: 2020-06-11"

## [1] "R version 4.0.0 (2020-04-24)"

### 1. Summary:

**The data set name (Grampian data) is misleading at the moment as it has all of Scotland’s SIMD data. Sorry**

This script adds additional information to the Scottish Index of Multiple Deprivation indicators (SIMD2020v2) data set. ***Consider adding more details e.g. best on openly accessible information bla and when information access. However, not 100% here would be the right place*** The current version of the script adds the postcodes (PC) of the respective SIMD2020v" data zones, their data zone names adn the NHS Health board regions.

### 2. Creating the data set

#### 2.1 Loading source data

setwd("~/grampian\_data/Raw\_data/")  
dir()

## [1] "NHS\_Health\_Board\_regions.csv" "SIMD2020v2datazones.csv"   
## [3] "SIMD2020v2indicators.csv" "SIMD2020v2postcodes.csv"

df\_SIMD2020.indi <- read.csv("SIMD2020v2indicators.csv")  
df\_SIMD2020.dz <- read.csv("SIMD2020v2datazones.csv")  
df\_SIMD2020.pc <- read.csv("SIMD2020v2postcodes.csv")  
df\_NHS\_regions <- read.csv("NHS\_Health\_Board\_regions.csv")

#### 2.2 Cleaning source data

##### 2.2.1 Removing excess data from source data

# From df\_SIMD.dz we only need the first two columns "DZ" & "DZname"  
df\_SIMD2020.dz <- df\_SIMD2020.dz[,c(1,2)]  
# From df\_SIMD2020.pc we only need the first two columns "Postcode" & "DZ"  
df\_SIMD2020.pc <- df\_SIMD2020.pc[,c(1,2)]

##### 2.2.2 Checking source data for duplicates

sapply(df\_SIMD2020.indi, function(x) sum(duplicated(x)))

## Data\_Zone Intermediate\_Zone Council\_area   
## 0 5726 6944   
## Total\_population Working\_age\_population Income\_rate   
## 6016 6201 6921   
## Income\_count Employment\_rate Employment\_count   
## 6605 6930 6766   
## CIF ALCOHOL DRUG   
## 6909 1109 2222   
## SMR DEPRESS LBWT   
## 1577 6936 6944   
## EMERG Attendance Attainment   
## 1685 6920 6684   
## no\_qualifications not\_participating University   
## 1119 6941 6926   
## drive\_petrol drive\_GP drive\_post   
## 5931 5996 6203   
## drive\_primary drive\_retail drive\_secondary   
## 6332 5682 5604   
## PT\_GP PT\_post PT\_retail   
## 4972 5278 4622   
## Broadband crime\_count crime\_rate   
## 6876 5636 927   
## overcrowded\_count nocentralheat\_count overcrowded\_rate   
## 6628 6856 6922   
## nocentralheat\_rate   
## 6955

# "Data\_Zone" should be 0, as they are our primary key here  
  
sapply(df\_SIMD2020.dz, function(x) sum(duplicated(x)))

## DZ DZname   
## 0 330

# "DZ" (Data zone) should be 0, as they are our primary key here  
  
sapply(df\_SIMD2020.pc, function(x) sum(duplicated(x)))

## Postcode DZ   
## 0 150858

# "Postcode" should be 0, as they are our primary key here

##### 2.2.3 Checking source data for NA-values

sapply(df\_SIMD2020.indi, function(x) sum(is.na(x)))

## Data\_Zone Intermediate\_Zone Council\_area   
## 0 0 0   
## Total\_population Working\_age\_population Income\_rate   
## 0 0 0   
## Income\_count Employment\_rate Employment\_count   
## 0 0 0   
## CIF ALCOHOL DRUG   
## 0 0 0   
## SMR DEPRESS LBWT   
## 0 0 0   
## EMERG Attendance Attainment   
## 0 0 0   
## no\_qualifications not\_participating University   
## 0 0 0   
## drive\_petrol drive\_GP drive\_post   
## 0 0 0   
## drive\_primary drive\_retail drive\_secondary   
## 0 0 0   
## PT\_GP PT\_post PT\_retail   
## 0 0 0   
## Broadband crime\_count crime\_rate   
## 0 0 0   
## overcrowded\_count nocentralheat\_count overcrowded\_rate   
## 0 0 0   
## nocentralheat\_rate   
## 0

# Data\_Zone should be 0. However, currently in the SIM2020v2 source data set missing values and suppressed values are denoted by "\*"   
  
# Lets check how many missing values and suppressed values are denoted by "\*" in the source data   
sapply(df\_SIMD2020.indi, function(x) sum(x=="\*"))

## Data\_Zone Intermediate\_Zone Council\_area   
## 0 0 0   
## Total\_population Working\_age\_population Income\_rate   
## 0 0 3   
## Income\_count Employment\_rate Employment\_count   
## 0 3 0   
## CIF ALCOHOL DRUG   
## 3 2 2   
## SMR DEPRESS LBWT   
## 2 1 1   
## EMERG Attendance Attainment   
## 2 567 189   
## no\_qualifications not\_participating University   
## 0 3 2   
## drive\_petrol drive\_GP drive\_post   
## 0 0 0   
## drive\_primary drive\_retail drive\_secondary   
## 0 0 0   
## PT\_GP PT\_post PT\_retail   
## 0 0 0   
## Broadband crime\_count crime\_rate   
## 2 500 501   
## overcrowded\_count nocentralheat\_count overcrowded\_rate   
## 0 0 0   
## nocentralheat\_rate   
## 0

sapply(df\_SIMD2020.dz, function(x) sum(is.na(x)))

## DZ DZname   
## 0 0

# DZ (Data zone) should be 0  
sapply(df\_SIMD2020.pc, function(x) sum(is.na(x)))

## Postcode DZ   
## 0 0

# Postcode should be 0

Nothing concerning here :)

### 2.3 Joining the source data set

#### 2.3.1 Merging/Joining data zone names to SIMD2020v2

df\_SIMD2020.1merge <- merge(df\_SIMD2020.indi, df\_SIMD2020.dz, by.x="Data\_Zone", by.y="DZ", all = TRUE) #all= TRUE to include potential missing values. In case something goes wrong with merge().  
  
# reordering the data frame, placing the DZname column in the second position  
df\_SIMD2020.1merge <- df\_SIMD2020.1merge[,c(1,38,2:37)]  
  
# Check if we introduced NA values  
sapply(df\_SIMD2020.1merge, function(x) sum(is.na(x)))

## Data\_Zone DZname Intermediate\_Zone   
## 0 0 0   
## Council\_area Total\_population Working\_age\_population   
## 0 0 0   
## Income\_rate Income\_count Employment\_rate   
## 0 0 0   
## Employment\_count CIF ALCOHOL   
## 0 0 0   
## DRUG SMR DEPRESS   
## 0 0 0   
## LBWT EMERG Attendance   
## 0 0 0   
## Attainment no\_qualifications not\_participating   
## 0 0 0   
## University drive\_petrol drive\_GP   
## 0 0 0   
## drive\_post drive\_primary drive\_retail   
## 0 0 0   
## drive\_secondary PT\_GP PT\_post   
## 0 0 0   
## PT\_retail Broadband crime\_count   
## 0 0 0   
## crime\_rate overcrowded\_count nocentralheat\_count   
## 0 0 0   
## overcrowded\_rate nocentralheat\_rate   
## 0 0

#### 2.3.2 Merging/Joining postcodes to SIMD2020v2

df\_SIMD2020.2merge <- merge(df\_SIMD2020.1merge, df\_SIMD2020.pc, by.x="Data\_Zone", by.y="DZ", all = TRUE)  
# reordering the data frame, placing the postcode column in the first position  
df\_SIMD2020.2merge <- df\_SIMD2020.2merge[,c(1,39,2:38)]  
  
  
# find the column with the NA values  
sapply(df\_SIMD2020.2merge, function(x) sum(is.na(x)))

## Data\_Zone Postcode DZname   
## 0 2 0   
## Intermediate\_Zone Council\_area Total\_population   
## 0 0 0   
## Working\_age\_population Income\_rate Income\_count   
## 0 0 0   
## Employment\_rate Employment\_count CIF   
## 0 0 0   
## ALCOHOL DRUG SMR   
## 0 0 0   
## DEPRESS LBWT EMERG   
## 0 0 0   
## Attendance Attainment no\_qualifications   
## 0 0 0   
## not\_participating University drive\_petrol   
## 0 0 0   
## drive\_GP drive\_post drive\_primary   
## 0 0 0   
## drive\_retail drive\_secondary PT\_GP   
## 0 0 0   
## PT\_post PT\_retail Broadband   
## 0 0 0   
## crime\_count crime\_rate overcrowded\_count   
## 0 0 0   
## nocentralheat\_count overcrowded\_rate nocentralheat\_rate   
## 0 0 0

# find affected rows  
df\_SIMD2020.2merge[is.na(df\_SIMD2020.2merge$Postcode),]

## Data\_Zone Postcode DZname Intermediate\_Zone Council\_area  
## 85669 S01010206 <NA> Petershill - 04 Petershill Glasgow City  
## 86063 S01010226 <NA> Sighthill - 02 Sighthill Glasgow City  
## Total\_population Working\_age\_population Income\_rate Income\_count  
## 85669 0 0 \* 0  
## 86063 0 0 \* 0  
## Employment\_rate Employment\_count CIF ALCOHOL DRUG SMR DEPRESS LBWT  
## 85669 \* 0 \* \* \* \* \* \*  
## 86063 \* 0 \* 95.22 57.20 153.32 0.01 0.00  
## EMERG Attendance Attainment no\_qualifications not\_participating  
## 85669 \* \* \* 353.08 \*  
## 86063 87.37 0.84 \* 202.42 0.00  
## University drive\_petrol drive\_GP drive\_post drive\_primary drive\_retail  
## 85669 \* 2.64 4.19 4.17 3.66 5.48  
## 86063 0.24 2.41 2.74 2.53 3.00 2.60  
## drive\_secondary PT\_GP PT\_post PT\_retail Broadband crime\_count crime\_rate  
## 85669 5.22 7.31 12.67 13.54 \* \* \*  
## 86063 2.92 7.93 10.40 9.51 \* \* \*  
## overcrowded\_count nocentralheat\_count overcrowded\_rate nocentralheat\_rate  
## 85669 243 21 0.49 0.04  
## 86063 339 45 0.42 0.06

There are no postcodes for those to data zone. I have contact [simd@scot.gov](mailto:simd@scot.gov) and currently wait for a reply. See issue #9 on [link](https://github.com/AbdnCHDS/grampian_data/issues).

## duplicated values  
sapply(df\_SIMD2020.2merge, function(x) sum(duplicated(x)))

## Data\_Zone Postcode DZname   
## 150858 1 151188   
## Intermediate\_Zone Council\_area Total\_population   
## 156584 157802 156874   
## Working\_age\_population Income\_rate Income\_count   
## 157059 157779 157463   
## Employment\_rate Employment\_count CIF   
## 157788 157624 157767   
## ALCOHOL DRUG SMR   
## 151967 153080 152435   
## DEPRESS LBWT EMERG   
## 157794 157802 152543   
## Attendance Attainment no\_qualifications   
## 157778 157542 151977   
## not\_participating University drive\_petrol   
## 157799 157784 156789   
## drive\_GP drive\_post drive\_primary   
## 156854 157061 157190   
## drive\_retail drive\_secondary PT\_GP   
## 156540 156462 155830   
## PT\_post PT\_retail Broadband   
## 156136 155480 157734   
## crime\_count crime\_rate overcrowded\_count   
## 156494 151785 157486   
## nocentralheat\_count overcrowded\_rate nocentralheat\_rate   
## 157714 157780 157813

# However, there is one duplicated postcode. The Postcodes should be 0  
# let find it  
  
df\_SIMD2020.2merge[duplicated(df\_SIMD2020.2merge$Postcode, fromLast=FALSE),]

## Data\_Zone Postcode DZname Intermediate\_Zone Council\_area  
## 86063 S01010226 <NA> Sighthill - 02 Sighthill Glasgow City  
## Total\_population Working\_age\_population Income\_rate Income\_count  
## 86063 0 0 \* 0  
## Employment\_rate Employment\_count CIF ALCOHOL DRUG SMR DEPRESS LBWT  
## 86063 \* 0 \* 95.22 57.20 153.32 0.01 0.00  
## EMERG Attendance Attainment no\_qualifications not\_participating  
## 86063 87.37 0.84 \* 202.42 0.00  
## University drive\_petrol drive\_GP drive\_post drive\_primary drive\_retail  
## 86063 0.24 2.41 2.74 2.53 3 2.6  
## drive\_secondary PT\_GP PT\_post PT\_retail Broadband crime\_count crime\_rate  
## 86063 2.92 7.93 10.4 9.51 \* \* \*  
## overcrowded\_count nocentralheat\_count overcrowded\_rate nocentralheat\_rate  
## 86063 339 45 0.42 0.06

# duplicated also picked up on the two NA in the postcodes

#### 2.3.2 Merging/Joining NHS Health Board regions to SIMD2020v2

df\_SIMD2020.3merge <- merge(df\_SIMD2020.2merge, df\_NHS\_regions, by.x="Council\_area", by.y="Council\_area", all = TRUE)  
# reordering the data frame  
df\_SIMD2020.3merge <- df\_SIMD2020.3merge[,c(2:5,1,40,6:39)]  
# find the column with the NA values  
sapply(df\_SIMD2020.3merge, function(x) sum(is.na(x)))

## Data\_Zone Postcode DZname   
## 0 2 0   
## Intermediate\_Zone Council\_area NHS\_Health\_Board\_Region   
## 0 0 0   
## Total\_population Working\_age\_population Income\_rate   
## 0 0 0   
## Income\_count Employment\_rate Employment\_count   
## 0 0 0   
## CIF ALCOHOL DRUG   
## 0 0 0   
## SMR DEPRESS LBWT   
## 0 0 0   
## EMERG Attendance Attainment   
## 0 0 0   
## no\_qualifications not\_participating University   
## 0 0 0   
## drive\_petrol drive\_GP drive\_post   
## 0 0 0   
## drive\_primary drive\_retail drive\_secondary   
## 0 0 0   
## PT\_GP PT\_post PT\_retail   
## 0 0 0   
## Broadband crime\_count crime\_rate   
## 0 0 0   
## overcrowded\_count nocentralheat\_count overcrowded\_rate   
## 0 0 0   
## nocentralheat\_rate   
## 0

# find affected rows  
df\_SIMD2020.3merge[is.na(df\_SIMD2020.3merge$Postcode),]

## Data\_Zone Postcode DZname Intermediate\_Zone Council\_area  
## 84706 S01010206 <NA> Petershill - 04 Petershill Glasgow City  
## 85100 S01010226 <NA> Sighthill - 02 Sighthill Glasgow City  
## NHS\_Health\_Board\_Region Total\_population Working\_age\_population  
## 84706 Greater Glasgow and Clyde 0 0  
## 85100 Greater Glasgow and Clyde 0 0  
## Income\_rate Income\_count Employment\_rate Employment\_count CIF ALCOHOL  
## 84706 \* 0 \* 0 \* \*  
## 85100 \* 0 \* 0 \* 95.22  
## DRUG SMR DEPRESS LBWT EMERG Attendance Attainment no\_qualifications  
## 84706 \* \* \* \* \* \* \* 353.08  
## 85100 57.20 153.32 0.01 0.00 87.37 0.84 \* 202.42  
## not\_participating University drive\_petrol drive\_GP drive\_post  
## 84706 \* \* 2.64 4.19 4.17  
## 85100 0.00 0.24 2.41 2.74 2.53  
## drive\_primary drive\_retail drive\_secondary PT\_GP PT\_post PT\_retail  
## 84706 3.66 5.48 5.22 7.31 12.67 13.54  
## 85100 3.00 2.60 2.92 7.93 10.40 9.51  
## Broadband crime\_count crime\_rate overcrowded\_count nocentralheat\_count  
## 84706 \* \* \* 243 21  
## 85100 \* \* \* 339 45  
## overcrowded\_rate nocentralheat\_rate  
## 84706 0.49 0.04  
## 85100 0.42 0.06

#### 2.4 Saving the data set

setwd("~/grampian\_data/Processed\_data/")  
  
write.csv(df\_SIMD2020.3merge, paste("Grampian\_data\_v",Dataset\_version,".csv", sep = ""), row.names=FALSE)