Remoteness-Data-Merge: R Code

###########################################################################  
# Script to collect, clean and merge remoteness data with unemployment data  
###########################################################################  
#  
# Assumptions   
#  
# 1) When rolling up SA1 area Remoteness type/rank to SA4 areas  
# the population denominator vale for each SA1 area is based on 2016  
# population levels for the entire 1999-2019 range of dates captured  
# within the unemployment dataset  
#  
# 2) Mapping of SA1 Remoteness areas to SA4 Unemployment levels is done based  
# on 2016 ASGS SA definitions and mapping and does not take into account any   
# possible changes to this mapping that might have taken place as part of the  
# 2011 ABS recalibration of ASGS areas  
#  
# 3) Incidence of 1,262 SA1 areas with 0 population does not materially impact  
# proof of concept of merge of Remoteness data to Unemployment data. Suggest   
# an invesitigation to ratify this assumption takes place before analysis is  
# advanced significantly  
  
library(readxl)  
library(tidyverse)  
library(raustats)  
  
# Read (from data) ABS excel download of 2016 Remoteness data  
RA\_2016\_AUST <- read\_excel("Data/RA\_2016\_AUST.xlsx")  
glimpse(RA\_2016\_AUST)  
  
# Read (from data) ABS excel download of 2011-19 population data  
POP\_AUST <- read\_excel("Data/POP\_AUST.xlsx")  
glimpse(POP\_AUST)  
  
# Read (from data) ABS SA1-to-SA4 csv data  
SA1\_4 <- read\_csv("Data/SA1\_2016\_AUST.csv")  
glimpse(SA1\_4)  
distinct(SA1\_4, SA4\_NAME\_2016)  
  
# Read (from data) ABS Uneployment data  
# Note: package tutorial in https://cran.r-project.org/web/packages/raustats/vignettes/raustats\_introduction.html   
load("data/unemployment.RData")  
  
# Tidy RA data  
RA <- RA\_2016\_AUST %>%  
 select(SA1\_7DIGITCODE\_2016, RA\_NAME\_2016) %>%  
 group\_by(SA1\_7DIGITCODE\_2016, RA\_NAME\_2016) %>%  
 summarize()  
glimpse(RA)  
  
# Tidy population data  
glimpse(POP\_AUST)  
is.na("POP\_AUST")  
# [1] FALSE  
test <- select(POP\_AUST, "2016")  
names(test) <- sub("2016", "Pop", names(test))  
any(test==0)  
#[1] TRUE  
test\_a <- arrange(test, Pop)  
glimpse(test\_a)  
count0 <- length(which(test\_a == 0))  
count0  
#[1] 1295  
# Not sure why 2016 contains 1,295 zero values - leave for now and reinvestigate   
# on refined data later  
  
POP <- select(POP\_AUST, SA1, "2016")  
names(POP) <- sub("^2016", "Pop", names(POP))  
glimpse(POP)  
  
  
# Tidy SA1-to-SA4 data  
SA1\_4 <- select(SA1\_4, SA1\_7DIGITCODE\_2016, SA4\_NAME\_2016) # Reduce the Number of columns  
glimpse(SA1\_4)  
  
  
# Add a handful of SA4 exeption names SA1\_4 so they'll match SA4 names in the Unemployment data  
amnd\_unemploy <- c("Greater Hobart", "New South Wales - Central West", "Victoria - North West", "Western Australia - Outback (North and South)", "Western Australia - Outback (North and South)", "Tasmania - South East", "Tasmania - West and North West")  
amnd\_sa4 <- c("Hobart", "Central West", "North West", "Western Australia - Outback (North)", "Western Australia - Outback (South)", "South East", "West and North West")  
for(i in 1:length(amnd\_sa4)){ SA1\_4$SA4\_NAME\_2016[SA1\_4$SA4\_NAME\_2016 == amnd\_sa4[i]] <- amnd\_unemploy[i] }  
filter(SA1\_4, SA4\_NAME\_2016 == "Greater Hobart"|SA4\_NAME\_2016 == "New South Wales - Central West"|SA4\_NAME\_2016 == "Victoria - North West"|SA4\_NAME\_2016 == "Western Australia - Outback (North and South)"|SA4\_NAME\_2016 == "Western Australia - Outback (North and South)"|SA4\_NAME\_2016 == "Tasmania - South East"|SA4\_NAME\_2016 == "Tasmania - West and North West") %>%  
 distinct(SA4\_NAME\_2016)  
  
#Tidy Unemployment - (already doen by virtue of Ana's code - no further coding necessary)  
  
# Merge SA1 Population data to SA1 Remoteness data  
SA1\_POP\_RA <- left\_join(POP, RA, by = c("SA1" = "SA1\_7DIGITCODE\_2016"))  
names(SA1\_POP\_RA) <- sub("RA\_NAME\_2016", "RA\_type", names(SA1\_POP\_RA))  
glimpse(SA1\_POP\_RA)  
  
# Add Remoteness Rank value to merged SA Pop. & Remoteness data  
SA1\_POP\_RA <- SA1\_POP\_RA %>% mutate(RA\_type\_rank = case\_when(  
 RA\_type == "Major Cities of Australia" ~ 1,  
 RA\_type == "Inner Regional Australia" ~ 2,   
 RA\_type == "Outer Regional Australia" ~ 3,   
 RA\_type == "Remote Australia" ~ 4,   
 RA\_type == "Very Remote Australia" ~ 5,  
 TRUE ~ 0)) %>%  
 filter(RA\_type\_rank != 0)  
glimpse(SA1\_POP\_RA)  
  
# Merge SA1 Population and Remoteness data to SA4 rollup levels  
SA1\_4\_POP\_RA <- left\_join(SA1\_POP\_RA, SA1\_4, by = c("SA1" = "SA1\_7DIGITCODE\_2016")) %>%  
 select(SA1, Pop, RA\_type, RA\_type\_rank, SA4\_NAME\_2016)  
names(SA1\_4\_POP\_RA) <- sub("SA4\_NAME\_2016", "SA4\_name", names(SA1\_4\_POP\_RA))  
glimpse(SA1\_4\_POP\_RA)  
SA1\_4\_POP\_RA %>% print(n = Inf)  
  
test\_a <- SA1\_4\_POP\_RA  
select(test\_a, SA4\_name, Pop)  
count0 <- length(which(test\_a == 0))  
count0  
# [1] 1262 zero values in "Pop" - why?  
test\_a <- SA1\_4\_POP\_RA  
select(test\_a, SA4\_name, RA\_type\_rank)  
count0 <- length(which(test\_a == 0))  
count0  
# [1] 1262 zero values in "RA\_Type\_rank" - why?  
test\_a <- select(SA1\_4\_POP\_RA, SA4\_name, Pop, RA\_type, RA\_type\_rank)  
test\_a <- filter(test\_a, Pop == 0)  
glimpse(test\_a)  
#Observations: 1,262  
#Variables: 4  
#$ SA4\_name <chr> "Capital Region", "Capital Region", "Capital Region", "Capital Region", "Capital Region...  
#$ Pop <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
#$ RA\_type <chr> "Major Cities of Australia", "Major Cities of Australia", "Major Cities of Australia", ...  
#$ RA\_type\_rank <dbl> 1, 1, 1, 1, 3, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, ...  
test\_a %>% print(n = 500)  
# Will need to access a file with SA1 names and merge with Pop data to resolve  
# Unanswered question of why there is so many 1,262 SA1 areas with 0 Pop. values  
  
#Add a population weighting factor to each SA1 to the Population and Remoteness SA1 and SA4 dataframe  
SA1\_4\_POP\_RA\_W <- SA1\_4\_POP\_RA %>%  
 group\_by(SA4\_name) %>%  
 summarize(SA4\_Pop = sum(Pop))  
SA1\_4\_POP\_RA\_W  
SA1\_4\_POP\_RA\_SumPop <- left\_join(SA1\_4\_POP\_RA, SA1\_4\_POP\_RA\_W)  
SA1\_4\_POP\_RA\_SumPop\_PopW <- mutate(SA1\_4\_POP\_RA\_SumPop, SA1\_Pop\_Wght = Pop / SA4\_Pop)  
SA1\_4\_POP\_RA\_SumPop\_PopW   
  
# Sample check that SA1\_Pop\_wght sum to 1 for an SA4\_Name  
glimpse(SA1\_4\_POP\_RA)  
SA1\_4\_POP\_RA\_test <- SA1\_4\_POP\_RA %>%  
 filter(SA4\_name == "Adelaide - Central and Hills") %>%  
 group\_by(SA4\_name) %>%  
 summarise(totL = sum(Pop))  
SA1\_4\_POP\_RA\_test  
# A tibble: 1 x 2  
#SA4\_name totL  
#<chr> <dbl>  
# 1 Adelaide - Central and Hills 297617  
# Checks out correctly  
  
SA1\_4\_POP\_RA\_SumPop\_PopW\_test <- SA1\_4\_POP\_RA\_SumPop\_PopW %>%  
 filter(SA4\_name == "Adelaide - Central and Hills") %>%  
 group\_by(SA4\_name) %>%  
 summarise(totL = sum(SA1\_Pop\_Wght))  
SA1\_4\_POP\_RA\_SumPop\_PopW\_test  
  
#Add a RA\_type\_rank\_Wghtd for each SA1 to the Population and Remoteness SA1 and SA4 dataframe  
SA1\_4\_POP\_RA\_SumPop\_PopW\_RATypeRankWghtd <- SA1\_4\_POP\_RA\_SumPop\_PopW %>%  
 mutate(RA\_type\_rank\_wghtd = RA\_type\_rank \* SA1\_Pop\_Wght)  
SA1\_4\_POP\_RA\_SumPop\_PopW\_RATypeRankWghtd  
  
# Construct final SA4\_name X PopWtdRA\_rank dataframe  
SA4\_RAPopWd <- SA1\_4\_POP\_RA\_SumPop\_PopW\_RATypeRankWghtd %>%  
 group\_by(SA4\_name) %>%  
 summarize(PopWtdRA\_rank = sum(RA\_type\_rank\_wghtd))  
SA4\_RAPopWd  
SA4\_RAPopWd %>% print(n = 100)  
  
#Join PopWtdRA\_rank to Unemployment data  
unemployment$territory\_sa4 = str\_trim(unemployment$territory\_sa4, side = "both")  
unemployment\_RAPopWtd <- left\_join(unemployment, SA4\_RAPopWd, by = c("territory\_sa4" = "SA4\_name"))  
glimpse(unemployment\_RAPopWtd)  
  
glimpse(unemployment\_RAPopWtd)  
distinct(unemployment\_RAPopWtd)  
  
save(unemployment\_RAPopWtd, file = "data/unemployment\_remote.RData")