PSTAT 231 Final Broderick-Weinberg

Hailey Broderick and Callum Weinberg

March 11, 2022

Executive Summary

Introduction

[Make reference early to the github page where this is hosted]

Background

Data

[DISCUSSION OF WHICH FOLDERS, CSVS, and VARIABLES are kept]

Purpose and Methods

Libraries

```
library(plyr)
library(dplyr)
library(tidyr)
library(assertr)
```

Functions

```
# From https://cran.r-project.org/web/packages/assertr/vignettes/assertr.html
not.missing.p <- function(x) if(is.na(x)) return(FALSE)
```

Data Cleaning

Manual Crosswalks

As mentioned above, within each folder from the Kaggle dataset, the differnt CSV files sometimes had the countries indexed to different values. In other words, there was not a unique identifier between the different CSV files for the Energy and Mineral folder and the Land and Agriculture folder. In order to combine (merge) these datasets, manual "crosswalks" had to be compiled, with a new identifier variable generated. These crosswalk files are available in the "Crosswalks" folder in the both the submission (note this applies to the submission for PSTAT 231) and the hosted github page (see introduction).

Additionally, there is a "final crosswalk" which maps the different subfolders (after the CSV from each sub folder has been combined) together is loaded.

```
## Energy and Minerals Crosswalk
energy_minerals_crosswalk_manual = read.csv("Crosswalks/Energy_Minerals_Crosswalk.csv")
energy_minerals_crosswalk_mineral_manual = read.csv("Crosswalks/Energy_Minerals_Crosswalk_mineral_datas

## Land and Agriculture Crosswalks
land_agriculture_preliminary_manual = read.csv("Crosswalks/land_agricultural_preliminary.csv")
land_agriculture_final_manual = read.csv("Crosswalks/land_agricultural_final.csv")

## Final Crosswalk for Mapping Categories
## Includes World Bank Data
final_crosswalk = read.csv("Crosswalks/02_final_crosswalk.csv")
final_crosswalk = final_crosswalk %>%
  rename(Country = Country_Mapping)
# Get version with only Country and ID
final_crosswalk_Country_Only = final_crosswalk %>%
  select(Country, Country_ID_Final)
```

Loading and Combing CSVs from Subfolders

The CSVs from each subfolder are loaded and cleaned. Cleaning in this case refers to keeping variables of interest, renaming variables, dropping empty rows, and checking that Country and Country ID are unique. For the folders that do have a unique country identifier for all CSVs in the folder, a crosswalk is generated in R and the CSVs are merged using that. Versions of the merged data are saved to the "intermediate" folder - this was done mainly to create the "final crosswalk" and is not strictly necessary if this code is being rerun.

Air and Climate Code:

```
## CH_4 Emissions
CH4_emissions = read.csv("Raw_Data/Air and Climate/CH4_Emissions.csv")

CH4_emissions = CH4_emissions %>%
  rename(Country_ID = Country.ID,
         CH4_latest_year = X.28,
         CH4 = CH4.emissions..latest.year,
         CH4_change_1990 = X..change.since.1990,
         CH4_per_capita = CH4.emissions..per.capita...latest.year) %>%
  select(Country_ID,Country,CH4_latest_year,CH4,CH4_change_1990,CH4_per_capita) %>%
  slice(2:n()) %>%
  assert(not.missing.p, Country_ID)

## CO2 Emissions
CO2_emissions = read.csv("Raw_Data/Air and Climate/CO2_Emissions.csv")

CO2_emissions = CO2_emissions %>%
  rename(Country_ID = Country.ID,
         CO2_latest_year = X.28,
         CO2 = CO2.emissions..latest.year,
         CO2_change_1990 = X..change.since.1990,
         CO2_per_capita = CO2.emissions..per.capita...latest.year) %>%
  select(Country_ID,Country,CO2_latest_year,CO2,CO2_change_1990,CO2_per_capita) %>%
  slice(2:n()) %>%
  assert(not.missing.p, Country_ID)

## GHG Emissions
```

```
GHG_emissions = read.csv("Raw_Data/Air and Climate/GHG_Emissions.csv")

GHG_emissions = GHG_emissions %>%
  rename(Country_ID = Country.ID,
         GHG_latest_year = X.28,
         GHG = GHG.total.without.LULUCF..latest.year,
         GHG_change_1990 = X..change.since.1990,
         GHG_per_capita = GHG.emissions.per.capita...latest.year) %>%
  select(Country_ID,Country,GHG_latest_year,GHG,GHG_change_1990,GHG_per_capita) %>%
  slice(2:n()) %>%
  assert(not.missing.p, Country_ID)

## GHG Emissions
GHG_emissions = read.csv("Raw_Data/Air and Climate/GHG_Emissions.csv")
# GHG_sector_total should be the same as GHG, if not investigate

GHG_emissions = GHG_emissions %>%
  rename(Country_ID = Country.ID,
         GHG_latest_year = X.28,
         GHG = GHG.total.without.LULUCF..latest.year,
         GHG_change_1990 = X..change.since.1990,
         GHG_per_capita = GHG.emissions.per.capita...latest.year) %>%
  select(Country_ID,Country,GHG_latest_year,GHG,GHG_change_1990,GHG_per_capita) %>%
  slice(2:n()) %>%
  assert(not.missing.p, Country_ID)

## GHG Emissions by Sector
# Add this in
# GHG_emissions_by_sector = read.csv("Data/Air and Climate/GHG_Emissions_by_Sector.csv")
# GHG_sector_total should be the same as GHG, if not investigate

## N2O Emissions
N2O_emissions = read.csv("Raw_Data/Air and Climate/N2O_Emissions.csv")

N2O_emissions = N2O_emissions %>%
  rename(Country_ID = Country.ID,
         N2O_latest_year = X.28,
         N2O = N2O.emissions..latest.year,
         N2O_change_1990 = X..change.since.1990,
         N2O_per_capita = N2O.emissions..per.capita...latest.year) %>%
  select(Country_ID,Country,N2O_latest_year,N2O,N2O_change_1990,N2O_per_capita) %>%
  slice(2:n()) %>%
  assert(not.missing.p, Country_ID)

## NOx Emissions
# There are some blank lines in this
# file that need to be skipped
NOx_emissions = read.csv("Raw_Data/Air and Climate/NOx_Emissions.csv")

NOx_emissions = NOx_emissions %>%
  rename(Country_ID = Country.ID,
         NOx_latest_year = X.28,
         NOx = NOx.emissions..latest.year,
```

```
         NOx_change_1990 = X..change.since.1990,
         NOx_per_capita = NOx..emissions.per.capita...latest.year) %>%
  select(Country_ID,Country,NOx_latest_year,NOx,NOx_change_1990,NOx_per_capita) %>%
  slice(2:173) %>%
  assert(not.missing.p, Country_ID)

## SO2 Emissions
SO2_emissions = read.csv("Raw_Data/Air and Climate/SO2_emissions.csv")

SO2_emissions = SO2_emissions %>%
  rename(Country_ID = Country.ID,
         SO2_latest_year = X.28,
         SO2 = SO2.emissions..latest.year,
         SO2_change_1990 = X..change.since.1990,
         SO2_per_capita = SO2.emissions.per.capita..latest.year) %>%
  select(Country_ID,Country,SO2_latest_year,SO2,SO2_change_1990,SO2_per_capita) %>%
  slice(2:143) %>%
  assert(not.missing.p, Country_ID)


# Append all of the Datasets
crosswalk_air_climate = rbind.fill(CH4_emissions,
                                   CO2_emissions,
                                   GHG_emissions,
                                   N2O_emissions,
                                   NOx_emissions,
                                   SO2_emissions)

# Create the Crosswalk
crosswalk_air_climate = crosswalk_air_climate %>%
  select(Country_ID, Country) %>%
  distinct()

# Check if there are any Duplicates
# This would mean that for the air and climate datasets
# there are either repeated countries, IDs, or that
# ID is not unique to country between the datasets
# and Vice versa
dim(crosswalk_air_climate[duplicated(crosswalk_air_climate$Country_ID),])[1] == 0
```

```
## [1] TRUE
```

```
dim(crosswalk_air_climate[duplicated(crosswalk_air_climate$Country),])[1] == 0
```

```
## [1] TRUE
```

```
# COmbine the datasets
# Start with the Above Created Crosswalk
# And Merge on Each of the Datasets
combined_air_climate =
  left_join(crosswalk_air_climate,CH4_emissions, by = "Country_ID") %>%
  select(-Country.y) %>%
  rename(Country = Country.x)
combined_air_climate =
  left_join(combined_air_climate,CO2_emissions, by = "Country_ID") %>%
```

```
  select(-Country.y) %>%
  rename(Country = Country.x)
combined_air_climate =
  left_join(combined_air_climate,GHG_emissions, by = "Country_ID") %>%
  select(-Country.y) %>%
  rename(Country = Country.x)
combined_air_climate =
  left_join(combined_air_climate,N2O_emissions, by = "Country_ID") %>%
  select(-Country.y) %>%
  rename(Country = Country.x)
combined_air_climate =
  left_join(combined_air_climate,NOx_emissions, by = "Country_ID") %>%
  select(-Country.y) %>%
  rename(Country = Country.x)
combined_air_climate =
  left_join(combined_air_climate,SO2_emissions, by = "Country_ID") %>%
  select(-Country.y) %>%
  rename(Country = Country.x)

# Export Data
save(combined_air_climate, file="Intermediate_Data/001_combined_air_climate.Rdata")
write.csv(combined_air_climate,"Intermediate_Data/001_combined_air_climate.csv", row.names = FALSE)

# Merge in Final Crosswalk to Get Universal ID
combined_air_climate_clean =
  left_join(final_crosswalk_Country_Only,combined_air_climate, by = "Country") %>%
  filter(!is.na(Country_ID)) %>%
  select(-Country, -Country_ID) %>%
  filter(Country_ID_Final != 999) #Excluded before merging

# Clean Up Environment
remove(crosswalk_air_climate,combined_air_climate, CH4_emissions, CO2_emissions,
       GHG_emissions,N2O_emissions, NOx_emissions, SO2_emissions)
```