Gathering Data

Ifechi Umeh

3/28/2022

*states vs countries. We need to define what a country is,* use entity instead of country

waste\_total <- read\_csv("Data Raw/plastic-waste-generation-total.csv")  
  
waste\_per\_cap <- read\_csv("Data Raw/plastic-waste-per-capita.csv")  
waste\_per\_cap <- waste\_per\_cap %>%   
 select(c("Entity", "Per capita plastic waste (kg/person/day)"))  
  
waste\_totgbl\_mismanage <- read\_csv("Data Raw/mismanaged-waste-global-total.csv")  
waste\_totgbl\_mismanage <- waste\_totgbl\_mismanage %>%   
 select(c("Entity", "Mismanaged waste (% global total)"))  
  
waste\_inadequate\_manage <- read\_csv("Data Raw/inadequately-managed-plastic.csv")  
waste\_inadequate\_manage <- waste\_inadequate\_manage %>%   
 select(c("Entity", "Share of plastic inadequately managed"))

\*Plastic data from Our World in Data

population\_ <- read\_csv("Data Raw/export.csv")  
pop <- read\_csv("Data Raw/export.csv")  
pop <- pop %>%   
 select(c("Country/Area Name", "Population", "Annual Growth Rate %", "Area (sq km)", "Density (per sq km)", "Total Fertility Rate", "Life Expectancy at Birth", "Under-5 Mortality Rate"))

Note: A lot of the N/a’s in the wiki webpage are smaller Islands nations/ territories. I would be a little uncomfortable with imputing this data since it is Informed Missing. I found that we could find some of those missing values

\*changed the population since Wikipedia is not a reliable source. Data from Census Bureau

bahamas

bahamas, the

country\_gdp <- read\_csv("Data Raw/Country-GPD-per-capita.csv")  
  
country\_gdp\_capita <- country\_gdp %>% transmute(Entity = `GDP per capita, current prices  
 (U.S. dollars per capita)`,  
 gdp\_per\_capita = `2010`)  
  
gdp\_per\_cap <- country\_gdp\_capita[c(2:197),] #to ensure we have just the countries and not regions

\*“GDP per capita, current prices (U.S. dollars per capita)”: Data from IMF

"Gross domestic product is the most commonly used single measure of a country’s overall economic activity. It represents the total value at current prices of final goods and services produced within a country during a specified time period divided by the average population for the same one year.

Source: World Economic Outlook (October 2021)" <https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD>

tab1 <- full\_join(waste\_total, pop, by = c("Entity" = "Country/Area Name"))  
tab2 <- full\_join(tab1, waste\_per\_cap, by = c("Entity"))  
tab3 <- full\_join(tab2, waste\_totgbl\_mismanage, by = c("Entity"))  
tab4 <- full\_join(tab3, waste\_inadequate\_manage, by = c("Entity"))  
tab5 <- full\_join(tab4, gdp\_per\_cap, by = c("Entity"))  
#tab6 <- left\_join(tab5, country\_gdp\_capita, by = "Entity") not sure what year is the USD in.

#write\_csv(tab5, "Data Clean/master\_table.csv")  
#write\_csv(tab5, "Data Clean/data\_1.csv")

*need to fix: some countries are spelt differently in different sources leading to duplicate rows that have NAs.* very difficult to find data on total plastic produced by country in 2010.

tab6 <- tab5  
# Tidying up Entity names  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Bahamas.\*", "Bahamas")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Brunei.\*", "Brunei")  
tab6$Entity <- str\_replace(tab6$Entity, "China, People's Republic of", "China")  
tab6$Entity <- str\_replace(tab6$Entity, "^Congo$", "Republic of Congo")  
tab6$Entity <- str\_replace(tab6$Entity, "Congo \\(Brazzaville\\)", "Republic of Congo")  
tab6$Entity <- str\_replace(tab6$Entity, "Congo, Republic of", "Republic of Congo")  
tab6$Entity <- str\_replace(tab6$Entity, "Congo \\(Kinshasa\\)", "Democratic Republic of Congo")  
tab6$Entity <- str\_replace(tab6$Entity, "Congo, Dem. Rep. of the", "Democratic Republic of Congo")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*C[ôo]te.\*", "Cote d'Ivoire")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Cura[cç]ao.\*", "Curacao")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Fa.?roe.\*", "Faeroe Islands")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Gambia.\*", "The Gambia")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Hong Kong.\*", "Hong Kong")  
tab6$Entity <- str\_replace(tab6$Entity, "Korea, North", "North Korea")  
tab6$Entity <- str\_replace(tab6$Entity, "Korea, South", "South Korea")  
tab6$Entity <- str\_replace(tab6$Entity, "Korea, Republic of", "South Korea")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Kyrgyz.\*", "Kyrgyzstan")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Lao.\*", "Laos")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Maca.\*", "Macao")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Micronesia.\*", "Micronesia")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Russia.\*", "Russia")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Saint Helena.\*", "Saint Helena")  
tab6$Entity <- str\_replace(tab6$Entity, "São Tomé and Príncipe", "Sao Tome and Principe")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Sint Maarten.\*", "Sint Maarten")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Slovak.\*", "Slovakia")  
tab6$Entity <- str\_replace(tab6$Entity, ".\*South Sudan.\*", "South Sudan") # South Sudan may need to be removed  
tab6$Entity <- str\_replace(tab6$Entity, ".\*Taiwan.\*", "Taiwan")  
tab6$Entity <- str\_replace(tab6$Entity, "Virgin Islands, British", "British Virgin Islands")  
  
  
tab7 <- tab6 %>%   
 group\_by(Entity) %>%   
 fill(everything(), .direction = "downup") %>%   
 slice(n=1)

* Congo Brazzaville (Republic of Congo) and Congo Kinshasa (Democratic Republic of Congo)
* country code: COG -> Republic of Congo; COD -> Democratic Republic of Congo *Note: South Sudan and Sudan were one country until they split in 2011. Issue because data is 2010. Possibly take out the South Sudan row* Not sure what to do about west bank and west bank and gaza

# NA EDA: Number of observations: 244  
sum(is.na(tab7$`Plastic waste generation (tonnes, total)`)) #Number of NAs: 76; proportion missing: 0.3114754  
sum(is.na(tab7$Population)) #Number of NAs: 17; proportion missing: 0.06967213  
sum(is.na(tab7$`Annual Growth Rate %`)) #Number of NAs: 18; proportion missing: 0.07377049 #US missing   
sum(is.na(tab7$`Area (sq km)`)) #Number of NAs: 17; proportion missing: 0.06967213  
sum(is.na(tab7$`Density (per sq km)`)) #Number of NAs: 17; proportion missing: 0.06967213  
sum(is.na(tab7$`Total Fertility Rate`)) #Number of NAs: 18; proportion missing: 0.07377049  
sum(is.na(tab7$`Life Expectancy at Birth`)) #Number of NAs: 18; proportion missing: 0.07377049  
sum(is.na(tab7$`Under-5 Mortality Rate`)) #Number of NAs: 18; proportion missing: 0.07377049  
sum(is.na(tab7$`Per capita plastic waste (kg/person/day)`)) #Number of NAs: 58; proportion missing: 0.2377049  
sum(is.na(tab7$`Mismanaged waste (% global total)`)) #Number of NAs: 58; proportion missing: 0.2377049  
sum(is.na(tab7$`Share of plastic inadequately managed`)) #Number of NAs: 58; proportion missing: 0.2377049  
sum(is.na(tab7$gdp\_per\_capita)) #Number of NAs: 48; proportion missing: 0.1967213

# Remove Code and Year  
tab8 <- select(tab7, -c("Code", "Year"))

write\_csv(tab8, "Data Clean/plastic\_1.csv")