# Install and load libraries

We need to first install all the required packages for the data analysis process. If you have never used the packages below, it is more likely that you have not installed them on your machine either. Please make sure you install each of the packages below using the following command:

install.packages("readxl")  
install.packages("here")  
install.packages("tidyverse")  
install.packages("janitor")  
install.packages("knitr")  
install.packages("flextable")  
install.packages("officer")  
install.packages("extrafont")  
# font\_import() # Only do this once

Then you need to load the following packages:

library(readxl)  
library(here)  
library(tidyverse)  
library(janitor)  
library(knitr)  
library(flextable)  
library(officer)  
library(extrafont)  
loadfonts(device = "win")

# Import clean dataset

There are two ways you can import the clean dataset into this file:

1. By running the R script from clean folder

source(here("scripts/cleaning", "unfccc-emissions-clean.R"))

This will also produce some of the temporary files.

1. By importing the clean dataset

unfccc\_emissions <- read.csv(here("scripts/cleaning",   
 "unfccc-emissions-clean.csv"))

It is better to load the cleaned dataset, because running the source script breaks down when it’s dealing with scripts. For example, the conversion of the script “Côte d’Ivoire” to “Cote d’Ivoire” for column-matching purposes does not translate properly when the script is run in the background. However, works well when it is done within the session and exported as a CSV sheet. Therefore, we recommend option 2 for the purpose of this exercise.

# Data exploration

We need to check for all the missing values, especially for Non-Annex I parties’ GHG inventories because they are voluntary submissions. First, we will check the degree of missing values for Annex I and Non-Annex I countries, starting with Non-Annex I countries.

table\_00 <- unfccc\_emissions %>%  
 dplyr::ungroup() %>%  
 dplyr::filter(group == "Non-Annex I") %>%  
 dplyr::group\_by(year) %>%  
 dplyr::summarise(missing = round(sum(is.na(co2))/length(co2) \*   
 100, digits = 2)) %>%  
 add\_row(year = "2019", missing = NA)  
  
table\_01 <- cbind(table\_00[table\_00$year <   
 2000, ], table\_00[table\_00$year >= 2000 &   
 table\_00$year < 2010, ], table\_00[table\_00$year >=   
 2010 & table\_00$year <= 2019, ]) %>%  
 repair\_names()  
  
autonum <- run\_autonum(seq\_id = "tab", bkm = "TC1",   
 bkm\_all = TRUE)  
  
table\_01 %>%  
 flextable::flextable() %>%  
 set\_header\_labels(year = "Year", missing = "Missing (%)",   
 year1 = "Year", missing1 = "Missing (%)",   
 year2 = "Year", missing2 = "Missing (%)") %>%  
 theme\_vanilla() %>%  
 font(fontname = "Times New Roman", part = "all") %>%  
 set\_caption("Percent of missing values of CO~2~ over time (Non-Annex I)",   
 autonum = autonum) %>%  
 highlight(j = "missing", i = ~missing <   
 50, color = "yellow") %>%  
 highlight(j = "missing1", i = ~missing1 <   
 50, color = "yellow") %>%  
 highlight(j = "missing2", i = ~missing2 <   
 50, color = "yellow") %>%  
 highlight(j = "missing", i = ~missing >   
 90, color = "lightcoral") %>%  
 highlight(j = "missing1", i = ~missing1 >   
 90, color = "lightcoral") %>%  
 highlight(j = "missing2", i = ~missing2 >   
 90, color = "lightcoral") %>%  
 autofit()

Table : Percent of missing values of CO2 over time (Non-Annex I)

| **Year** | **Missing (%)** | **Year** | **Missing (%)** | **Year** | **Missing (%)** |
| --- | --- | --- | --- | --- | --- |
| 1990 | 68.92 | 2000 | 29.39 | 2010 | 71.62 |
| 1991 | 87.16 | 2001 | 84.46 | 2011 | 89.19 |
| 1992 | 85.81 | 2002 | 82.43 | 2012 | 83.11 |
| 1993 | 83.78 | 2003 | 82.43 | 2013 | 87.84 |
| 1994 | 28.38 | 2004 | 81.08 | 2014 | 92.57 |
| 1995 | 78.38 | 2005 | 71.62 | 2015 | 93.24 |
| 1996 | 83.78 | 2006 | 78.38 | 2016 | 94.59 |
| 1997 | 83.11 | 2007 | 83.11 | 2017 | 97.97 |
| 1998 | 81.76 | 2008 | 84.46 | 2018 | 99.32 |
| 1999 | 85.14 | 2009 | 85.81 | 2019 |  |

As we can see the data deprivations are quite stark when it comes to reporting data on GHG and CO2 emissions for most developing countries. There are only two years where data reporting exceeds 50% of the official members: 1994, and 2000.

table\_00 <- unfccc\_emissions %>%  
 dplyr::ungroup() %>%  
 dplyr::filter(group == "Annex I") %>%  
 dplyr::group\_by(year) %>%  
 dplyr::summarise(missing = round(sum(is.na(co2))/length(co2) \*   
 100, digits = 2)) %>%  
 add\_row(year = "2019", missing = NA)  
  
table\_01 <- cbind(table\_00[table\_00$year <   
 2000, ], table\_00[table\_00$year >= 2000 &   
 table\_00$year < 2010, ], table\_00[table\_00$year >=   
 2010 & table\_00$year <= 2019, ]) %>%  
 repair\_names()  
  
  
table\_01 %>%  
 flextable::flextable() %>%  
 set\_header\_labels(year = "Year", missing = "Missing (%)",   
 year1 = "Year", missing1 = "Missing (%)",   
 year2 = "Year", missing2 = "Missing (%)") %>%  
 font(fontname = "Times New Roman", part = "all") %>%  
 set\_caption("Percent of missing values of CO~2~ over time (Annex I)",   
 autonum = autonum) %>%  
 theme\_vanilla() %>%  
 autofit()

Table : Percent of missing values of CO2 over time (Annex I)

| **Year** | **Missing (%)** | **Year** | **Missing (%)** | **Year** | **Missing (%)** |
| --- | --- | --- | --- | --- | --- |
| 1990 | 0 | 2000 | 0 | 2010 | 0 |
| 1991 | 0 | 2001 | 0 | 2011 | 0 |
| 1992 | 0 | 2002 | 0 | 2012 | 0 |
| 1993 | 0 | 2003 | 0 | 2013 | 0 |
| 1994 | 0 | 2004 | 0 | 2014 | 0 |
| 1995 | 0 | 2005 | 0 | 2015 | 0 |
| 1996 | 0 | 2006 | 0 | 2016 | 0 |
| 1997 | 0 | 2007 | 0 | 2017 | 0 |
| 1998 | 0 | 2008 | 0 | 2018 | 0 |
| 1999 | 0 | 2009 | 0 | 2019 |  |

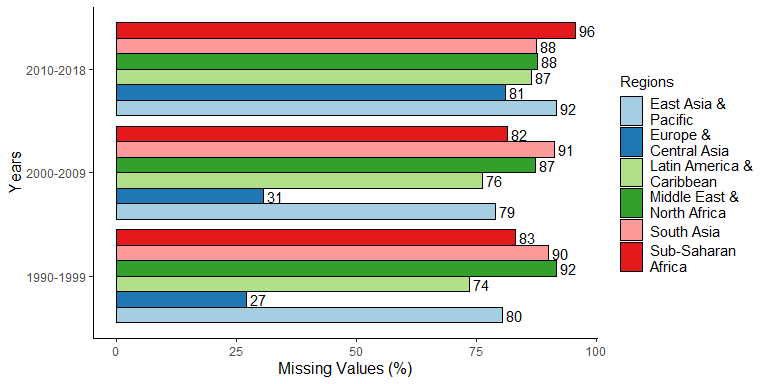
The story is not the same for Annex I countries because reporting GHG and CO2 emission inventories is mandatory within the UNFCCC.

Next, we should check how data deprivations vary by region within the Non-Annex I countries, that way we can determine which region to focus on for increasing data capacity.

First we need to aggregate the data by region, and we will split them by three time periods: 1990-1999, 2000-2009, and 2010-2018.

missing\_non\_annex\_i\_region <- unfccc\_emissions %>%  
 dplyr::ungroup() %>%  
 dplyr::filter(group == "Non-Annex I") %>%  
 dplyr::mutate(time\_period = floor(as.numeric(as.character(year))/10) \*   
 10) %>%  
 dplyr::group\_by(time\_period, region) %>%  
 dplyr::summarise(missing = sum(is.na(co2))/length(co2) \*   
 100) %>%  
 dplyr::mutate(time\_period = as.factor(time\_period))  
  
missing\_non\_annex\_i\_region$time\_period <- dplyr::recode\_factor(missing\_non\_annex\_i\_region$time\_period,   
 `1990` = "1990-1999", `2000` = "2000-2009",   
 `2010` = "2010-2018")

missing\_non\_annex\_i\_region %>%  
 ggplot2::ggplot(aes(x = time\_period,   
 y = missing, fill = str\_wrap(region,   
 15))) + ggplot2::geom\_bar(stat = "identity",   
 position = position\_dodge(), color = "black") +   
 ggplot2::theme\_classic() + ggplot2::ylab("Missing Values (%)") +   
 ggplot2::xlab("Years") + ggplot2::coord\_flip() +   
 ggplot2::scale\_fill\_brewer(palette = "Paired",   
 name = "Regions") + ggplot2::geom\_text(aes(label = round(missing,   
 digits = 0)), position = position\_dodge(width = 0.9),   
 hjust = -0.25, size = 4) + ggplot2::theme(text = element\_text(family = "Arial",   
 size = 11), axis.title.x = element\_text(family = "Arial",   
 size = 12), axis.title.y = element\_text(family = "Arial",   
 size = 12), legend.text = element\_text(family = "Arial",   
 size = 11))



ggsave(here("images", "missing-non-annex-i-01.svg"),   
 device = "svg", dpi = 300)

We can see that many Non-Annex I party submit their GHG inventories the least which makes it difficult to aggregate their cumulative emissions, with the highest levels of missing values located in Africa and Oceania. We will use data from the Global Carbon Project to calculate cumulative CO2 based on their estimates.

## Export as an R script for future use

Only run this chunk manually once within the .Rmd file. It produces an error when knitting it as a whole because of chunk label duplicates. As of May 10, 2021, there hasn’t been a viable solution to run the code below when as part of the knitting process.

knitr::purl("unfccc-emissions-analysis.Rmd",   
 "unfccc-emissions-analysis.R")  
knitr::write\_bib(.packages(), "packages.bib")

# Software used

Arel-Bundock, Vincent. *Countrycode: Convert Country Names and Country Codes*, 2020. <https://github.com/vincentarelbundock/countrycode>.

Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman. “Countrycode: An r Package to Convert Country Names and Country Codes.” *Journal of Open Source Software* 3, no. 28 (2018): 848. <https://doi.org/10.21105/joss.00848>.

Firke, Sam. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*, 2021. <https://github.com/sfirke/janitor>.

Gohel, David. *Flextable: Functions for Tabular Reporting*, 2021. <https://CRAN.R-project.org/package=flextable>.

———. *Officer: Manipulation of Microsoft Word and PowerPoint Documents*, 2021. <https://CRAN.R-project.org/package=officer>.

Henry, Lionel, and Hadley Wickham. *Purrr: Functional Programming Tools*, 2020. <https://CRAN.R-project.org/package=purrr>.

Müller, Kirill. *Here: A Simpler Way to Find Your Files*, 2020. <https://CRAN.R-project.org/package=here>.

Müller, Kirill, and Hadley Wickham. *Tibble: Simple Data Frames*, 2021. <https://CRAN.R-project.org/package=tibble>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. <https://www.R-project.org/>.

Robinson, David. *Fuzzyjoin: Join Tables Together on Inexact Matching*, 2020. <https://github.com/dgrtwo/fuzzyjoin>.

Schauberger, Philipp, and Alexander Walker. *Openxlsx: Read, Write and Edit Xlsx Files*, 2020. <https://CRAN.R-project.org/package=openxlsx>.

Wickham, Hadley. *Forcats: Tools for Working with Categorical Variables (Factors)*, 2021. <https://CRAN.R-project.org/package=forcats>.

———. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. <https://ggplot2.tidyverse.org>.

———. *Stringr: Simple, Consistent Wrappers for Common String Operations*, 2019. <https://CRAN.R-project.org/package=stringr>.

———. *Tidyr: Tidy Messy Data*, 2021. <https://CRAN.R-project.org/package=tidyr>.

———. *Tidyverse: Easily Install and Load the Tidyverse*, 2021. <https://CRAN.R-project.org/package=tidyverse>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. “Welcome to the tidyverse.” *Journal of Open Source Software* 4, no. 43 (2019): 1686. <https://doi.org/10.21105/joss.01686>.

Wickham, Hadley, and Jennifer Bryan. *Readxl: Read Excel Files*, 2019. <https://CRAN.R-project.org/package=readxl>.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2020. <https://CRAN.R-project.org/package=ggplot2>.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. *Dplyr: A Grammar of Data Manipulation*, 2021. <https://CRAN.R-project.org/package=dplyr>.

Wickham, Hadley, and Jim Hester. *Readr: Read Rectangular Text Data*, 2020. <https://CRAN.R-project.org/package=readr>.

Winston Chang. *Extrafont: Tools for Using Fonts*, 2014. <https://github.com/wch/extrafont>.

Xie, Yihui. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC, 2015. <https://yihui.org/knitr/>.

———. “Knitr: A Comprehensive Tool for Reproducible Research in R.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC, 2014. <http://www.crcpress.com/product/isbn/9781466561595>.

———. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*, 2021. <https://yihui.org/knitr/>.