Data Consolidation and Cleaning

## Data Cleaning and Consolidation

## There were 2 rounds of data cleaning and consolidation involved. First round happened when each members have prepared and cleaned each individual ##datasets for each domain. After the cleaned dataset was passed on to other group members, some follow-up for data cleaning and consolidation was ##done.

# 1. Read Libraries

library("readxl")  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.2  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'stringr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)

# 2. Load all datasets

Education

education\_latest\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Education\_Datasets\_Summary.xlsx", sheet = "Latest\_Available\_Data"))  
  
education\_5YearsMean\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Education\_Datasets\_Summary.xlsx", sheet = "Past\_5\_Years\_Mean"))

Health

health\_latest\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/HealthData\_MergedNew.xlsx", sheet = "LatestAvailableData"))  
  
health\_5YearsMean\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/HealthData\_MergedNew.xlsx", sheet = "Past5YrsMean"))

Health Facilities - Data Type Conversion was made to some variables here

# Health Facilities  
  
health\_fac\_latest\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/LW\_Datasets\_Merged.xlsx", sheet = "Latest"))  
  
health\_fac\_5YearsMean\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/LW\_Datasets\_Merged.xlsx", sheet = "Mean"))  
  
  
  
# Transform the chr datatype into numeric  
  
# latest data  
  
health\_fac\_latest\_data\_convert <- health\_fac\_latest\_data %>%   
 select(-c('Country\_Name', 'Country\_Code')) %>%   
 mutate\_if(is.character, as.numeric)

## Warning: Problem with `mutate()` input `Nurse and Midwives(per1,000People)`.  
## i NAs introduced by coercion  
## i Input `Nurse and Midwives(per1,000People)` is `.Primitive("as.double")(`Nurse and Midwives(per1,000People)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Physician(per1,000People)`.  
## i NAs introduced by coercion  
## i Input `Physician(per1,000People)` is `.Primitive("as.double")(`Physician(per1,000People)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `OECD\_Nurse(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `OECD\_Nurse(1000HAB)` is `.Primitive("as.double")(`OECD\_Nurse(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `OECD\_Doctor(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `OECD\_Doctor(1000HAB)` is `.Primitive("as.double")(`OECD\_Doctor(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Health Expenditure (% of GDP)`.  
## i NAs introduced by coercion  
## i Input `Health Expenditure (% of GDP)` is `.Primitive("as.double")(`Health Expenditure (% of GDP)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Hospital\_Bed\_ACUTE(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `Hospital\_Bed\_ACUTE(1000HAB)` is `.Primitive("as.double")(`Hospital\_Bed\_ACUTE(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Hospital\_Bed\_ICU(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `Hospital\_Bed\_ICU(1000HAB)` is `.Primitive("as.double")(`Hospital\_Bed\_ICU(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Hospital\_Bed\_OTHER(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `Hospital\_Bed\_OTHER(1000HAB)` is `.Primitive("as.double")(`Hospital\_Bed\_OTHER(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` is `.Primitive("as.double")(`HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `HOSPITAL\_BED\_TOTAL(1000HAB)`.  
## i NAs introduced by coercion  
## i Input `HOSPITAL\_BED\_TOTAL(1000HAB)` is `.Primitive("as.double")(`HOSPITAL\_BED\_TOTAL(1000HAB)`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Health Care Index \_2000`.  
## i NAs introduced by coercion  
## i Input `Health Care Index \_2000` is `.Primitive("as.double")(`Health Care Index \_2000`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `HealthCare\_Rank\_2000`.  
## i NAs introduced by coercion  
## i Input `HealthCare\_Rank\_2000` is `.Primitive("as.double")(HealthCare\_Rank\_2000)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Covid19\_test\_conducted\_people tested`.  
## i NAs introduced by coercion  
## i Input `Covid19\_test\_conducted\_people tested` is `.Primitive("as.double")(`Covid19\_test\_conducted\_people tested`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Covid19\_test\_conducted\_samples tested`.  
## i NAs introduced by coercion  
## i Input `Covid19\_test\_conducted\_samples tested` is `.Primitive("as.double")(`Covid19\_test\_conducted\_samples tested`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Covid19\_test\_conducted\_tests performed`.  
## i NAs introduced by coercion  
## i Input `Covid19\_test\_conducted\_tests performed` is `.Primitive("as.double")(`Covid19\_test\_conducted\_tests performed`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Covid19\_test\_conducted\_units unclear`.  
## i NAs introduced by coercion  
## i Input `Covid19\_test\_conducted\_units unclear` is `.Primitive("as.double")(`Covid19\_test\_conducted\_units unclear`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

glimpse(health\_fac\_latest\_data\_convert)

## Rows: 308  
## Columns: 16  
## $ `Nurse and Midwives(per1,000People)` <dbl> NA, 0.175500, 0.407500, NA...  
## $ `Physician(per1,000People)` <dbl> NA, 0.278200, 0.214600, NA...  
## $ `OECD\_Nurse(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `OECD\_Doctor(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Health Expenditure (% of GDP)` <dbl> NA, 11.777194, 2.791500, N...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Health Care Index \_2000` <dbl> NA, 0.325, 0.275, NA, NA, ...  
## $ HealthCare\_Rank\_2000 <dbl> NA, 173, 181, NA, NA, 55, ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> NA, NA, NA, NA, NA, NA, NA...

health\_fac\_latest\_data <- cbind(health\_fac\_latest\_data[,c('Country\_Name', 'Country\_Code')], health\_fac\_latest\_data\_convert)  
  
  
glimpse(health\_fac\_latest\_data)

## Rows: 308  
## Columns: 18  
## $ Country\_Name <chr> "ARUBA", "AFGHANISTAN", "A...  
## $ Country\_Code <chr> "ABW", "AFG", "AGO", "AIA"...  
## $ `Nurse and Midwives(per1,000People)` <dbl> NA, 0.175500, 0.407500, NA...  
## $ `Physician(per1,000People)` <dbl> NA, 0.278200, 0.214600, NA...  
## $ `OECD\_Nurse(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `OECD\_Doctor(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Health Expenditure (% of GDP)` <dbl> NA, 11.777194, 2.791500, N...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Health Care Index \_2000` <dbl> NA, 0.325, 0.275, NA, NA, ...  
## $ HealthCare\_Rank\_2000 <dbl> NA, 173, 181, NA, NA, 55, ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> NA, NA, NA, NA, NA, NA, NA...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> NA, NA, NA, NA, NA, NA, NA...

# 5 years mean  
  
health\_fac\_5YearsMean\_data\_convert <- health\_fac\_5YearsMean\_data %>%   
 select(-c('Country\_Name', 'Country\_Code')) %>%   
 mutate\_if(is.character, as.numeric)

## Warning: Problem with `mutate()` input `Nurse and Midwives(per1,000People)\_2016\_2019`.  
## i NAs introduced by coercion  
## i Input `Nurse and Midwives(per1,000People)\_2016\_2019` is `.Primitive("as.double")(`Nurse and Midwives(per1,000People)\_2016\_2019`)`.  
  
## Warning: NAs introduced by coercion

## Warning: Problem with `mutate()` input `Physician(per1,000People)\_2016\_2019`.  
## i NAs introduced by coercion  
## i Input `Physician(per1,000People)\_2016\_2019` is `.Primitive("as.double")(`Physician(per1,000People)\_2016\_2019`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `OECD\_Nurse(1000HAB)\_2016\_2019`.  
## i NAs introduced by coercion  
## i Input `OECD\_Nurse(1000HAB)\_2016\_2019` is `.Primitive("as.double")(`OECD\_Nurse(1000HAB)\_2016\_2019`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `OECD\_Doctor(1000HAB)\_2016\_2019`.  
## i NAs introduced by coercion  
## i Input `OECD\_Doctor(1000HAB)\_2016\_2019` is `.Primitive("as.double")(`OECD\_Doctor(1000HAB)\_2016\_2019`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

## Warning: Problem with `mutate()` input `Health Expenditure (% of GDP)\_2016\_2017`.  
## i NAs introduced by coercion  
## i Input `Health Expenditure (% of GDP)\_2016\_2017` is `.Primitive("as.double")(`Health Expenditure (% of GDP)\_2016\_2017`)`.

## Warning in mask$eval\_all\_mutate(dots[[i]]): NAs introduced by coercion

glimpse(health\_fac\_5YearsMean\_data\_convert)

## Rows: 308  
## Columns: 5  
## $ `Nurse and Midwives(per1,000People)\_2016\_2019` <dbl> NA, 0.1618500, 0.407...  
## $ `Physician(per1,000People)\_2016\_2019` <dbl> NA, 0.2782000, 0.214...  
## $ `OECD\_Nurse(1000HAB)\_2016\_2019` <dbl> NA, NA, NA, NA, NA, ...  
## $ `OECD\_Doctor(1000HAB)\_2016\_2019` <dbl> NA, NA, NA, NA, NA, ...  
## $ `Health Expenditure (% of GDP)\_2016\_2017` <dbl> NA, 11.369589, 2.752...

health\_fac\_5YearsMean\_data <- cbind(health\_fac\_5YearsMean\_data[,c('Country\_Name', 'Country\_Code')], health\_fac\_5YearsMean\_data\_convert)  
  
  
glimpse(health\_fac\_5YearsMean\_data)

## Rows: 308  
## Columns: 7  
## $ Country\_Name <chr> "ARUBA", "AFGHANISTA...  
## $ Country\_Code <chr> "ABW", "AFG", "AGO",...  
## $ `Nurse and Midwives(per1,000People)\_2016\_2019` <dbl> NA, 0.1618500, 0.407...  
## $ `Physician(per1,000People)\_2016\_2019` <dbl> NA, 0.2782000, 0.214...  
## $ `OECD\_Nurse(1000HAB)\_2016\_2019` <dbl> NA, NA, NA, NA, NA, ...  
## $ `OECD\_Doctor(1000HAB)\_2016\_2019` <dbl> NA, NA, NA, NA, NA, ...  
## $ `Health Expenditure (% of GDP)\_2016\_2017` <dbl> NA, 11.369589, 2.752...

General Data

general\_data\_latest <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Ricky-Data-Summary.xlsx", sheet = "Latest"))  
general\_data\_5YearsMean\_data <- as.data.frame(read\_excel("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Ricky-Data-Summary.xlsx", sheet = "5\_Year\_Mean"))

Covid Data

covid\_data <- read.csv("https://raw.githubusercontent.com/datasets/covid-19/master/data/countries-aggregated.csv")  
  
latest\_covid <- filter(covid\_data,Date == '2020-12-19') #Latest Available Data  
  
# view first 5 rows of covid data  
head(latest\_covid)

## Date Country Confirmed Recovered Deaths  
## 1 2020-12-19 Afghanistan 49681 38613 2047  
## 2 2020-12-19 Albania 52542 27831 1074  
## 3 2020-12-19 Algeria 94781 62869 2659  
## 4 2020-12-19 Andorra 7560 6963 80  
## 5 2020-12-19 Angola 16626 9518 386  
## 6 2020-12-19 Antigua and Barbuda 152 141 5

# 3.Joining / Merging tables

5 Year Mean Data

### Combining Latest Covid Date with 5-Year Mean Data ###  
  
  
# 1. Left Join Education to the covid data  
  
mean\_data\_1 <- left\_join(latest\_covid, education\_5YearsMean\_data, by = c("Country" = "Country Name"))   
  
# 2. Left join Health to 1.  
  
mean\_data\_2 <- left\_join(mean\_data\_1, health\_5YearsMean\_data, by = c("Country" = "Country Name"))   
  
  
# 3. Left join General Data to 2.  
  
mean\_data\_3 <- left\_join(mean\_data\_2, general\_data\_5YearsMean\_data, by = c("Country" = "Country Name"))   
  
  
# 4. Left join General Data to 3.  
  
mean\_data\_4 <- left\_join(mean\_data\_3, health\_fac\_5YearsMean\_data, by = c("Country Code.x" = "Country\_Code"))   
  
  
# Drop repeated & Unnecessary Columns  
  
glimpse(mean\_data\_4)

## Rows: 198  
## Columns: 40  
## $ Date <fct> ...  
## $ Country <chr> ...  
## $ Confirmed <int> ...  
## $ Recovered <int> ...  
## $ Deaths <int> ...  
## $ `Country Code.x` <chr> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Country Code.y` <chr> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ Country\_Name <chr> ...  
## $ `Nurse and Midwives(per1,000People)\_2016\_2019` <dbl> ...  
## $ `Physician(per1,000People)\_2016\_2019` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)\_2016\_2019` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)\_2016\_2019` <dbl> ...  
## $ `Health Expenditure (% of GDP)\_2016\_2017` <dbl> ...

mean\_data\_dropped\_redundant\_cols <- mean\_data\_4 %>% select(-c('Country Code.x', 'Country Code.y', 'Country\_Name'))  
  
  
glimpse(mean\_data\_dropped\_redundant\_cols)

## Rows: 198  
## Columns: 37  
## $ Date <fct> ...  
## $ Country <chr> ...  
## $ Confirmed <int> ...  
## $ Recovered <int> ...  
## $ Deaths <int> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Nurse and Midwives(per1,000People)\_2016\_2019` <dbl> ...  
## $ `Physician(per1,000People)\_2016\_2019` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)\_2016\_2019` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)\_2016\_2019` <dbl> ...  
## $ `Health Expenditure (% of GDP)\_2016\_2017` <dbl> ...

# Check missing values of each variables (columns)  
  
mean\_data\_view\_missing\_data <- colSums(is.na(mean\_data\_dropped\_redundant\_cols) == TRUE)  
  
glimpse(mean\_data\_view\_missing\_data)

## Named num [1:37] 0 0 0 0 0 25 25 25 25 25 ...  
## - attr(\*, "names")= chr [1:37] "Date" "Country" "Confirmed" "Recovered" ...

5 Year Latest Data (2019)

### Combining Latest Covid Date with the latest Data ###  
  
  
# 1. Left Join Education to the covid data  
  
latest\_data\_1 <- left\_join(latest\_covid, education\_latest\_data, by = c("Country" = "Country Name"))   
  
  
# 2. Left join Health to 1.  
  
latest\_data\_2 <- left\_join(latest\_data\_1, health\_latest\_data, by = c("Country" = "Country Name"))   
  
  
  
# 3. Left join General Data to 2.  
  
latest\_data\_3 <- left\_join(latest\_data\_2, general\_data\_latest, by = c("Country" = "Country Name"))   
  
  
  
# 4. Left join General Data to 3.  
  
latest\_data\_4 <- left\_join(latest\_data\_3, health\_fac\_latest\_data, by = c("Country Code.x" = "Country\_Code"))   
  
  
  
# Drop repeated & Unnecessary Columns  
  
glimpse(latest\_data\_4)

## Rows: 198  
## Columns: 63  
## $ Date <fct> ...  
## $ Country <chr> ...  
## $ Confirmed <int> ...  
## $ Recovered <int> ...  
## $ Deaths <int> ...  
## $ `Country Code.x` <chr> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Country Code.y` <chr> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Country Code` <chr> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Net ODA Received (2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `Inflation (2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `GDP (2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Unemployment (2019)` <dbl> ...  
## $ `Happiness Score (2019)` <dbl> ...  
## $ `GDP per capita (2019)` <dbl> ...  
## $ `Social support (2019)` <dbl> ...  
## $ `Healthy life expectancy (2019)` <dbl> ...  
## $ `Freedom to make life choices (2019)` <dbl> ...  
## $ `Generosity (2019)` <dbl> ...  
## $ `Perceptions of corruption (2019)` <dbl> ...  
## $ Country\_Name <chr> ...  
## $ `Nurse and Midwives(per1,000People)` <dbl> ...  
## $ `Physician(per1,000People)` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)` <dbl> ...  
## $ `Health Expenditure (% of GDP)` <dbl> ...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> ...  
## $ `Health Care Index \_2000` <dbl> ...  
## $ HealthCare\_Rank\_2000 <dbl> ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> ...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> ...

latest\_data\_dropped\_redundant\_cols <- latest\_data\_4 %>% select(-c('Country Code.x', 'Country Code.y', 'Country Code', 'Country\_Name'))  
  
  
glimpse(latest\_data\_dropped\_redundant\_cols)

## Rows: 198  
## Columns: 59  
## $ Date <fct> ...  
## $ Country <chr> ...  
## $ Confirmed <int> ...  
## $ Recovered <int> ...  
## $ Deaths <int> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Net ODA Received (2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `Inflation (2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `GDP (2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Unemployment (2019)` <dbl> ...  
## $ `Happiness Score (2019)` <dbl> ...  
## $ `GDP per capita (2019)` <dbl> ...  
## $ `Social support (2019)` <dbl> ...  
## $ `Healthy life expectancy (2019)` <dbl> ...  
## $ `Freedom to make life choices (2019)` <dbl> ...  
## $ `Generosity (2019)` <dbl> ...  
## $ `Perceptions of corruption (2019)` <dbl> ...  
## $ `Nurse and Midwives(per1,000People)` <dbl> ...  
## $ `Physician(per1,000People)` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)` <dbl> ...  
## $ `Health Expenditure (% of GDP)` <dbl> ...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> ...  
## $ `Health Care Index \_2000` <dbl> ...  
## $ HealthCare\_Rank\_2000 <dbl> ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> ...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> ...

# Check missing values of each variables (columns)  
  
latest\_data\_view\_missing\_data <- colSums(is.na(latest\_data\_dropped\_redundant\_cols) == TRUE)  
  
glimpse(latest\_data\_view\_missing\_data)

## Named num [1:59] 0 0 0 0 0 25 25 25 25 25 ...  
## - attr(\*, "names")= chr [1:59] "Date" "Country" "Confirmed" "Recovered" ...

# Findings: There are more data for LATEST DATASET, hence the analysis will be done using latest dataset

glimpse(latest\_data\_dropped\_redundant\_cols)

## Rows: 198  
## Columns: 59  
## $ Date <fct> ...  
## $ Country <chr> ...  
## $ Confirmed <int> ...  
## $ Recovered <int> ...  
## $ Deaths <int> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Net ODA Received (2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `Inflation (2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `GDP (2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Unemployment (2019)` <dbl> ...  
## $ `Happiness Score (2019)` <dbl> ...  
## $ `GDP per capita (2019)` <dbl> ...  
## $ `Social support (2019)` <dbl> ...  
## $ `Healthy life expectancy (2019)` <dbl> ...  
## $ `Freedom to make life choices (2019)` <dbl> ...  
## $ `Generosity (2019)` <dbl> ...  
## $ `Perceptions of corruption (2019)` <dbl> ...  
## $ `Nurse and Midwives(per1,000People)` <dbl> ...  
## $ `Physician(per1,000People)` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)` <dbl> ...  
## $ `Health Expenditure (% of GDP)` <dbl> ...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> ...  
## $ `Health Care Index \_2000` <dbl> ...  
## $ HealthCare\_Rank\_2000 <dbl> ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> ...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> ...

analysis\_dataset <- latest\_data\_dropped\_redundant\_cols

# 4. Save Dataset

### Save ###  
  
library("writexl")

## Warning: package 'writexl' was built under R version 3.6.3

write\_xlsx(analysis\_dataset,"consolidate\_dataset.xlsx")

# Follow Up - 2nd round of cleaning

# Merge latest dataset

latest\_data\_combined <- merge(x = education\_latest\_data, y = health\_latest\_data, by = "Country Code")  
  
latest\_data\_combined2 <- merge(x = latest\_data\_combined, y = general\_data\_latest, by = "Country Code")  
  
latest\_data\_combined3 <- merge(x = latest\_data\_combined2, y = health\_fac\_latest\_data, by.x = "Country Code", by.y = "Country\_Code")  
  
glimpse(latest\_data\_combined3)

## Rows: 226  
## Columns: 59  
## $ `Country Code` <chr> ...  
## $ `Country Name.x` <chr> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Country Name.y` <chr> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Country Name` <chr> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Net ODA Received (2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `Inflation (2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `GDP (2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Unemployment (2019)` <dbl> ...  
## $ `Happiness Score (2019)` <dbl> ...  
## $ `GDP per capita (2019)` <dbl> ...  
## $ `Social support (2019)` <dbl> ...  
## $ `Healthy life expectancy (2019)` <dbl> ...  
## $ `Freedom to make life choices (2019)` <dbl> ...  
## $ `Generosity (2019)` <dbl> ...  
## $ `Perceptions of corruption (2019)` <dbl> ...  
## $ Country\_Name <chr> ...  
## $ `Nurse and Midwives(per1,000People)` <dbl> ...  
## $ `Physician(per1,000People)` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)` <dbl> ...  
## $ `Health Expenditure (% of GDP)` <dbl> ...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> ...  
## $ `Health Care Index \_2000` <dbl> ...  
## $ HealthCare\_Rank\_2000 <dbl> ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> ...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> ...

# Drop redundant data

latest\_data\_combined4 <- latest\_data\_combined3 %>% select(-c("Country Name.x", "Country Name.y", "Country\_Name", "Country Name"))  
  
names(latest\_data\_combined4)[names(latest\_data\_combined4) == 'Country Code'] <- "Country\_Code"  
  
  
glimpse(latest\_data\_combined4)

## Rows: 226  
## Columns: 55  
## $ Country\_Code <chr> ...  
## $ `School enrollment, tertiary (% gross)` <dbl> ...  
## $ `School enrollment, secondary (% net)` <dbl> ...  
## $ `Literacy rate, adult total (% of people ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult female (% of females ages 15 and above)` <dbl> ...  
## $ `Literacy rate, adult male (% of males ages 15 and above)` <dbl> ...  
## $ `Literacy rate, youth male (% of males ages 15-24)` <dbl> ...  
## $ `Literacy rate, youth female (% of females ages 15-24)` <dbl> ...  
## $ `Labor force, female (% of total labor force)` <dbl> ...  
## $ `Labor force, total` <dbl> ...  
## $ `Population ages 15-64 (% of total population)` <dbl> ...  
## $ `Government expenditure on education, total (% of GDP)` <dbl> ...  
## $ `Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)` <dbl> ...  
## $ `Cause of death, by injury (% of total)` <dbl> ...  
## $ `Cause of death, by non-communicable diseases (% of total)` <dbl> ...  
## $ `Death rate, crude (per 1,000 people)` <dbl> ...  
## $ `Diabetes prevalence (% of population ages 20 to 79)` <dbl> ...  
## $ `Immunization, DPT (% of children ages 12-23 months)` <dbl> ...  
## $ `Immunization, HepB3 (% of one-year-old children)` <dbl> ...  
## $ `Immunization, measles (% of children ages 12-23 months)` <dbl> ...  
## $ `Life expectancy at birth, total (years)` <dbl> ...  
## $ `Prevalence of anemia among children (% of children under 5)` <dbl> ...  
## $ `Prevalence of undernourishment (% of population)` <dbl> ...  
## $ `Risk of impoverishing expenditure for surgical care (% of people at risk)` <dbl> ...  
## $ `Net ODA Received (2014-2018)` <dbl> ...  
## $ `Net ODA Received (2018)` <dbl> ...  
## $ `Inflation (2015-2019)` <dbl> ...  
## $ `Inflation (2019)` <dbl> ...  
## $ `GDP (2015-2019)` <dbl> ...  
## $ `GDP (2019)` <dbl> ...  
## $ `Unemployment (2015-2019)` <dbl> ...  
## $ `Unemployment (2019)` <dbl> ...  
## $ `Happiness Score (2019)` <dbl> ...  
## $ `GDP per capita (2019)` <dbl> ...  
## $ `Social support (2019)` <dbl> ...  
## $ `Healthy life expectancy (2019)` <dbl> ...  
## $ `Freedom to make life choices (2019)` <dbl> ...  
## $ `Generosity (2019)` <dbl> ...  
## $ `Perceptions of corruption (2019)` <dbl> ...  
## $ `Nurse and Midwives(per1,000People)` <dbl> ...  
## $ `Physician(per1,000People)` <dbl> ...  
## $ `OECD\_Nurse(1000HAB)` <dbl> ...  
## $ `OECD\_Doctor(1000HAB)` <dbl> ...  
## $ `Health Expenditure (% of GDP)` <dbl> ...  
## $ `Hospital\_Bed\_ACUTE(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_ICU(1000HAB)` <dbl> ...  
## $ `Hospital\_Bed\_OTHER(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_PSYCHIATRIC(1000HAB)` <dbl> ...  
## $ `HOSPITAL\_BED\_TOTAL(1000HAB)` <dbl> ...  
## $ `Health Care Index \_2000` <dbl> ...  
## $ HealthCare\_Rank\_2000 <dbl> ...  
## $ `Covid19\_test\_conducted\_people tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_samples tested` <dbl> ...  
## $ `Covid19\_test\_conducted\_tests performed` <dbl> ...  
## $ `Covid19\_test\_conducted\_units unclear` <dbl> ...

# Read continents data

continent <- read.csv("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets//Cleaning/continents2.csv")  
  
head(continent)

## ï..name alpha\_2 alpha\_3 country\_code iso\_3166\_2 region  
## 1 Afghanistan AF AFG 4 ISO 3166-2:AF Asia  
## 2 Ã…land Islands AX ALA 248 ISO 3166-2:AX Europe  
## 3 Albania AL ALB 8 ISO 3166-2:AL Europe  
## 4 Algeria DZ DZA 12 ISO 3166-2:DZ Africa  
## 5 American Samoa AS ASM 16 ISO 3166-2:AS Oceania  
## 6 Andorra AD AND 20 ISO 3166-2:AD Europe  
## sub\_region intermediate\_region region\_code sub\_region\_code  
## 1 Southern Asia 142 34  
## 2 Northern Europe 150 154  
## 3 Southern Europe 150 39  
## 4 Northern Africa 2 15  
## 5 Polynesia 9 61  
## 6 Southern Europe 150 39  
## intermediate\_region\_code  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

continent2 <- continent %>% select(c('ï..name', alpha\_3, region, sub\_region, intermediate\_region))  
   
 names(continent2)[names(continent2) == "ï..name"] <- "Country\_Name"  
 names(continent2)[names(continent2) == "alpha\_3"] <- "Country\_Code"  
 names(continent2)[names(continent2) == "region"] <- "Region"  
 names(continent2)[names(continent2) == "sub\_region"] <- "Sub\_Region"  
 names(continent2)[names(continent2) == "intermediate\_region"] <- "Intermediate\_Region"  
  
  
  
glimpse(latest\_covid)

## Rows: 191  
## Columns: 5  
## $ Date <fct> 2020-12-19, 2020-12-19, 2020-12-19, 2020-12-19, 2020-12-1...  
## $ Country <fct> Afghanistan, Albania, Algeria, Andorra, Angola, Antigua a...  
## $ Confirmed <int> 49681, 52542, 94781, 7560, 16626, 152, 1537169, 153173, 2...  
## $ Recovered <int> 38613, 27831, 62869, 6963, 9518, 141, 1362617, 131931, 25...  
## $ Deaths <int> 2047, 1074, 2659, 80, 386, 5, 41763, 2616, 908, 5209, 217...

glimpse(continent2)

## Rows: 249  
## Columns: 5  
## $ Country\_Name <fct> Afghanistan, Ã…land Islands, Albania, Algeria, ...  
## $ Country\_Code <fct> AFG, ALA, ALB, DZA, ASM, AND, AGO, AIA, ATA, AT...  
## $ Region <fct> Asia, Europe, Europe, Africa, Oceania, Europe, ...  
## $ Sub\_Region <fct> Southern Asia, Northern Europe, Southern Europe...  
## $ Intermediate\_Region <fct> , , , , , , Middle Africa, Caribbean, , Caribbe...

# Merge Continent and Covid Data

glimpse(latest\_covid)

## Rows: 191  
## Columns: 5  
## $ Date <fct> 2020-12-19, 2020-12-19, 2020-12-19, 2020-12-19, 2020-12-1...  
## $ Country <fct> Afghanistan, Albania, Algeria, Andorra, Angola, Antigua a...  
## $ Confirmed <int> 49681, 52542, 94781, 7560, 16626, 152, 1537169, 153173, 2...  
## $ Recovered <int> 38613, 27831, 62869, 6963, 9518, 141, 1362617, 131931, 25...  
## $ Deaths <int> 2047, 1074, 2659, 80, 386, 5, 41763, 2616, 908, 5209, 217...

glimpse(continent2)

## Rows: 249  
## Columns: 5  
## $ Country\_Name <fct> Afghanistan, Ã…land Islands, Albania, Algeria, ...  
## $ Country\_Code <fct> AFG, ALA, ALB, DZA, ASM, AND, AGO, AIA, ATA, AT...  
## $ Region <fct> Asia, Europe, Europe, Africa, Oceania, Europe, ...  
## $ Sub\_Region <fct> Southern Asia, Northern Europe, Southern Europe...  
## $ Intermediate\_Region <fct> , , , , , , Middle Africa, Caribbean, , Caribbe...

DFCovid2 <- merge(x = continent2, y = latest\_covid, by.x = "Country\_Name", by.y = "Country",all.y = TRUE) # right join  
  
  
#remove date column  
DFCovid3 <- DFCovid2 %>% select(-c(Date))  
  
  
glimpse(DFCovid3)

## Rows: 191  
## Columns: 8  
## $ Country\_Name <fct> Afghanistan, Albania, Algeria, Andorra, Angola,...  
## $ Country\_Code <fct> AFG, ALB, DZA, AND, AGO, ATG, ARG, ARM, AUS, AU...  
## $ Region <fct> Asia, Europe, Africa, Europe, Africa, Americas,...  
## $ Sub\_Region <fct> Southern Asia, Southern Europe, Northern Africa...  
## $ Intermediate\_Region <fct> , , , , Middle Africa, Caribbean, South America...  
## $ Confirmed <int> 49681, 52542, 94781, 7560, 16626, 152, 1537169,...  
## $ Recovered <int> 38613, 27831, 62869, 6963, 9518, 141, 1362617, ...  
## $ Deaths <int> 2047, 1074, 2659, 80, 386, 5, 41763, 2616, 908,...

# Join df with aggregated data and Manual Preprocessing for US country name

# import samuel data  
agg\_data <- read.csv("C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Cleaning/consolidate\_dataset\_edited\_FilteredCSV3.csv")  
  
agg\_data2 <- agg\_data %>% select(c(code,ConfirmedPerPopulation...,RecoveryRate...,DeathRate...))  
  
glimpse(agg\_data2)

## Rows: 185  
## Columns: 4  
## $ code <fct> AFG, AGO, ALB, AND, ARE, ARG, ARM, ATG, A...  
## $ ConfirmedPerPopulation... <dbl> 0.131, 0.052, 1.841, 9.800, 1.969, 3.421,...  
## $ RecoveryRate... <dbl> 77.722, 57.248, 52.969, 92.103, 87.383, 8...  
## $ DeathRate... <dbl> 4.120, 2.322, 2.044, 1.058, 0.330, 2.717,...

names(agg\_data2)[names(agg\_data2) == "code"] <- "Country\_Code"  
names(agg\_data2)[names(agg\_data2) == "ConfirmedPerPopulation..."] <- "ConfirmedPerPpopulation"  
names(agg\_data2)[names(agg\_data2) == "RecoveryRate..."] <- "RecoveryRate"  
names(agg\_data2)[names(agg\_data2) == "DeathRate..."] <- "DeathRate"  
  
  
glimpse(agg\_data2)

## Rows: 185  
## Columns: 4  
## $ Country\_Code <fct> AFG, AGO, ALB, AND, ARE, ARG, ARM, ATG, AUS...  
## $ ConfirmedPerPpopulation <dbl> 0.131, 0.052, 1.841, 9.800, 1.969, 3.421, 5...  
## $ RecoveryRate <dbl> 77.722, 57.248, 52.969, 92.103, 87.383, 88....  
## $ DeathRate <dbl> 4.120, 2.322, 2.044, 1.058, 0.330, 2.717, 1...

Covid\_Rate\_Data <- merge(x = DFCovid3, y = agg\_data2, by.x = "Country\_Code", by.y = "Country\_Code",all.x = TRUE)  
  
# Change the Country\_Name , US to United States  
  
Covid\_Rate\_Data[193, 2] <- 'United States'  
Covid\_Rate\_Data[193, 1] <- 'USA'  
  
head(Covid\_Rate\_Data)

## Country\_Code Country\_Name Region Sub\_Region  
## 1 AFG Afghanistan Asia Southern Asia  
## 2 AGO Angola Africa Sub-Saharan Africa  
## 3 ALB Albania Europe Southern Europe  
## 4 AND Andorra Europe Southern Europe  
## 5 ARE United Arab Emirates Asia Western Asia  
## 6 ARG Argentina Americas Latin America and the Caribbean  
## Intermediate\_Region Confirmed Recovered Deaths ConfirmedPerPpopulation  
## 1 49681 38613 2047 0.131  
## 2 Middle Africa 16626 9518 386 0.052  
## 3 52542 27831 1074 1.841  
## 4 7560 6963 80 9.800  
## 5 192404 168129 634 1.969  
## 6 South America 1537169 1362617 41763 3.421  
## RecoveryRate DeathRate  
## 1 77.722 4.120  
## 2 57.248 2.322  
## 3 52.969 2.044  
## 4 92.103 1.058  
## 5 87.383 0.330  
## 6 88.645 2.717

# Combine the latest data with covid data and write data into csv

glimpse(Covid\_Rate\_Data)

## Rows: 198  
## Columns: 11  
## $ Country\_Code <fct> AFG, AGO, ALB, AND, ARE, ARG, ARM, ATG, AUS...  
## $ Country\_Name <fct> Afghanistan, Angola, Albania, Andorra, Unit...  
## $ Region <fct> Asia, Africa, Europe, Europe, Asia, America...  
## $ Sub\_Region <fct> Southern Asia, Sub-Saharan Africa, Southern...  
## $ Intermediate\_Region <fct> , Middle Africa, , , , South America, , Car...  
## $ Confirmed <int> 49681, 16626, 52542, 7560, 192404, 1537169,...  
## $ Recovered <int> 38613, 9518, 27831, 6963, 168129, 1362617, ...  
## $ Deaths <int> 2047, 386, 1074, 80, 634, 41763, 2616, 5, 9...  
## $ ConfirmedPerPpopulation <dbl> 0.131, 0.052, 1.841, 9.800, 1.969, 3.421, 5...  
## $ RecoveryRate <dbl> 77.722, 57.248, 52.969, 92.103, 87.383, 88....  
## $ DeathRate <dbl> 4.120, 2.322, 2.044, 1.058, 0.330, 2.717, 1...

latest\_data <- merge(x = Covid\_Rate\_Data, y = latest\_data\_combined4, by.x = "Country\_Code", by.y = "Country\_Code",all.x = TRUE)  
  
  
  
latest\_data\_ver2 <- unique(latest\_data)  
  
  
# check how many repeated countries  
  
unique\_country <- unique(latest\_data$Country\_Code)  
dim(unique\_country)

## NULL

unique\_country2 <- unique(latest\_data$Country\_Name)  
dim(unique\_country2)

## NULL

write.csv(latest\_data\_ver2,"C:/Users/uwxl003/Downloads/PoDS - Consolidate Datasets/PoDS - Consolidate Datasets/Cleaning/latest\_cleansed\_data.csv", row.names = FALSE)