creating_gap_data

Dav King

2/23/2022

Note: gapminder does not enable an aggregate downloading of their data. Thus, I have built out a dataset here. My apologies for not iterating my development on github, but given that this was not technically a part of the assignment, I don't think it's a major issue. I have left definitions for each variable in comments where they are created. Ultimately the dataset is built and stored as an Rdata file named `gap`.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
men_15_24 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men_15
men_25_34 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men_25
men_35_44 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men_35
men_45_54 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men_45
men_55_64 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men_55
men_65_plus <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_men
women_15_24 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_wome
women_25_34 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_wome
women_35_44 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_wome
women_45_54 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_wome
women_55_64 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_wome
women_65_plus <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school_w

ed <- function(edData){
  edData %>%
    pivot_longer(
      cols = !ï..country,
    names_to = "year",
    values_to = "ed"
    )
}
```

```
men_15_24 <- ed(men_15_24)
men_25_34 <- ed(men_25_34)
men_35_44 <- ed(men_35_44)
men_45_54 <- ed(men_45_54)
men_55_64 <- ed(men_55_64)
men_65_plus <- ed(men_65_plus)
women_15_24 <- ed(women_15_24)
women_25_34 <- ed(women_25_34)
women_35_44 <- ed(women_35_44)
women_45_54 <- ed(women_45_54)
women_55_64 <- ed(women_55_64)
women_65_plus <- ed(women_65_plus)

#All consider mean years of schooling, separated by age group and sex
educ <- men_15_24 %>%
  full_join(men_25_34, by = c("ï..country", "year")) %>%
  full_join(men_35_44, by = c("ï..country", "year")) %>%
  full_join(men_45_54, by = c("ï..country", "year")) %>%
  full_join(men_55_64, by = c("ï..country", "year")) %>%
  full_join(men_65_plus, by = c("ï..country", "year")) %>%
  full_join(women_15_24, by = c("ï..country", "year")) %>%
  full_join(women_25_34, by = c("ï..country", "year")) %>%
  full_join(women_35_44, by = c("ï..country", "year")) %>%
  full_join(women_45_54, by = c("ï..country", "year")) %>%
  full_join(women_55_64, by = c("ï..country", "year")) %>%
  full_join(women_65_plus, by = c("ï..country", "year"))

names(educ) <- c("country", "year", "ed_men_15_24", "ed_men_25_34",
                 "ed_men_35_44", "ed_men_45_54", "ed_men_55_64",
                 "ed_men_65_plus", "ed_women_15_24", "ed_women_25_34",
                 "ed_women_35_44", "ed_women_45_54",
                 "ed_women_55_64", "ed_women_65_plus")
```

```
hdi <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/hdi_human_development_index.csv")
#Human development index score
hdi <- hdi %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "hdi"
  )

life_exp_birth <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/life_expectancy_years
#Mean years expected at birth
life_exp_birth <- life_exp_birth %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "life_exp_birth"
  )

gnipercap_ppp <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/gnipercapita_ppp_curre
#GNI Per Capita, PPP, current international $
```

```r
gnipercap_ppp <- gnipercap_ppp %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "gnipercap_ppp"
  )
```

```r
co2_emissions <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/co2_emissions_tonnes_p
#CO2 emissions, metric tonnes per person
co2_emissions <- co2_emissions %>%
  select(-X1829, -X1830, -X1831, -X1832) %>%
  #note: this fixes issue with non-numeric values
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "co2_emissions"
  )

child_mortality <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/child_mortality_0_5_
#Deaths of children under 5 years per 1,000 live births
child_mortality <- child_mortality %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "child_mortality"
  )

pop <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/population_total.csv")
#Total Population
#Note: this absolutely foul data will need serious transformation before it can
#be used, idk what Gapminder was thinking
pop <- pop %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "pop"
  )

gini <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/gini.csv")
#gini coefficient of wealth inequality
gini <- gini %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "gini"
  )

poverty <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/number_of_people_in_poverty.
#Number of poor population, in millions, living on less than $1.25/day at 2005
#international prices
poverty <- poverty %>%
  pivot_longer(
    cols = !ï..country,
```

```
    names_to = "year",
    values_to = "poverty"
  )

cell_phones_per_100 <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/cell_phones_per_
#Cell phones per 100 people
cell_phones_per_100$X1986[36] <- NA
cell_phones_per_100$X1986 <- as.double(cell_phones_per_100$X1986)
#Note: remove one non-numeric value
cell_phones_per_100 <- cell_phones_per_100 %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "cell_phones_per_100"
  )

pct_not_using_internet <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/non_net_users
#Percentage of people not using the internet in the last 3 months
pct_not_using_internet <- pct_not_using_internet %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "pct_not_using_internet"
  )

journalists_killed <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/journakilled.csv"
#Number of journalists killed in a year
journalists_killed <- journalists_killed %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "journalists_killed"
  )

ed_gender_ratio <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/mean_years_in_school
#ratio of female to male number of years in school, 25-34 year olds
ed_gender_ratio <- ed_gender_ratio %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "ed_gender_ratio"
  )

literacy_rate <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/literacy_rate_adult_to
#Literacy rate, age 15+
literacy_rate <- literacy_rate %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "literacy_rate"
  )

primary_school_comp <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/primary_completi
```

```r
#Primary school completion rate
primary_school_comp <- primary_school_comp %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "primary_school_comp"
  )

primary_student_spending <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/expenditure
#Government expenditure per primary school student, as a percentage of GDP
#per capita
primary_student_spending <- primary_student_spending %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "primary_student_spending"
  )
```

```r
antivax <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/vccin_sfty_dag.csv")
#Percentage of people who disagree that vaccines are safe for children to have
antivax <- antivax %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "antivax"
  )

health_spending <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/total_health_spendin
#Average health expenditure per person, in USD, using average exchange rate
health_spending <- health_spending %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "health_spending"
  )
```

```r
median_age <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/median_age_years.csv")
#Median age of the total population, in years
median_age <- median_age %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "median_age"
  )

pop_dens <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/population_density_per_squa
#People per square kilometer
pop_dens <- pop_dens %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "pop_dens"
  )
```

```r
urban_pop <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/urban_population_percent_o
#Percentage of people living in urban areas (defined by national statistical
#offices)
urban_pop <- urban_pop %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "urban_pop"
  )

urban_pop_growth <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/urban_population_gr
#Percentage of people living in urban areas, annual % growth
urban_pop_growth <- urban_pop_growth %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "urban_pop_growth"
  )

first_marriage_age <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/age_at_1st_marria
#Mean age of first marriage for women
first_marriage_age <- first_marriage_age %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "first_marriage_age"
  )

babies_per_woman <- read.csv("~/Stat 199/project-spring-2022-group-2-8-project/data/children_per_woman_
#Fertility rate: the number of children that would be born to each woman with
#prevailing age-specific fertility rates
babies_per_woman <- babies_per_woman %>%
  pivot_longer(
    cols = !ï..country,
    names_to = "year",
    values_to = "babies_per_woman"
  )
```

```r
gap <- hdi %>%
  full_join(pop, by = c("ï..country", "year")) %>%
  full_join(median_age, by = c("ï..country", "year")) %>%
  full_join(life_exp_birth, by = c("ï..country", "year")) %>%
  full_join(pop_dens, by = c("ï..country", "year")) %>%
  full_join(educ, by = c("ï..country" = "country", "year")) %>%
  full_join(gnipercap_ppp, by = c("ï..country", "year")) %>%
  full_join(co2_emissions, by = c("ï..country", "year")) %>%
  full_join(child_mortality, by = c("ï..country", "year")) %>%
  full_join(gini, by = c("ï..country", "year")) %>%
  full_join(poverty, by = c("ï..country", "year")) %>%
  full_join(cell_phones_per_100, by = c("ï..country", "year")) %>%
  full_join(pct_not_using_internet, by = c("ï..country", "year")) %>%
  full_join(journalists_killed, by = c("ï..country", "year")) %>%
  full_join(ed_gender_ratio, by = c("ï..country", "year")) %>%
```

```
  full_join(literacy_rate, by = c("ï..country", "year")) %>%
  full_join(primary_school_comp, by = c("ï..country", "year")) %>%
  full_join(primary_student_spending, by = c("ï..country", "year")) %>%
  full_join(antivax, by = c("ï..country", "year")) %>%
  full_join(health_spending, by = c("ï..country", "year")) %>%
  full_join(urban_pop, by = c("ï..country", "year")) %>%
  full_join(urban_pop_growth, by = c("ï..country", "year")) %>%
  full_join(first_marriage_age, by = c("ï..country", "year")) %>%
  full_join(babies_per_woman, by = c("ï..country", "year")) %>%
  arrange(ï..country, year)
gap
```

```
## # A tibble: 100,616 x 37
##    ï..country  year    hdi pop   median_age life_exp_birth pop_dens ed_men_15_24
##    <chr>       <chr> <dbl> <chr>      <dbl>          <dbl> <chr>           <dbl>
##  1 Afghanistan X1615    NA <NA>          NA             NA <NA>               NA
##  2 Afghanistan X1616    NA <NA>          NA             NA <NA>               NA
##  3 Afghanistan X1617    NA <NA>          NA             NA <NA>               NA
##  4 Afghanistan X1618    NA <NA>          NA             NA <NA>               NA
##  5 Afghanistan X1619    NA <NA>          NA             NA <NA>               NA
##  6 Afghanistan X1620    NA <NA>          NA             NA <NA>               NA
##  7 Afghanistan X1621    NA <NA>          NA             NA <NA>               NA
##  8 Afghanistan X1622    NA <NA>          NA             NA <NA>               NA
##  9 Afghanistan X1623    NA <NA>          NA             NA <NA>               NA
## 10 Afghanistan X1624    NA <NA>          NA             NA <NA>               NA
## # ... with 100,606 more rows, and 29 more variables: ed_men_25_34 <dbl>,
## #   ed_men_35_44 <dbl>, ed_men_45_54 <dbl>, ed_men_55_64 <dbl>,
## #   ed_men_65_plus <dbl>, ed_women_15_24 <dbl>, ed_women_25_34 <dbl>,
## #   ed_women_35_44 <dbl>, ed_women_45_54 <dbl>, ed_women_55_64 <dbl>,
## #   ed_women_65_plus <dbl>, gnipercap_ppp <chr>, co2_emissions <dbl>,
## #   child_mortality <dbl>, gini <dbl>, poverty <dbl>,
## #   cell_phones_per_100 <dbl>, pct_not_using_internet <dbl>, ...
```

```
save(gap, file = "gap.Rdata")
```