Creating data set

William John Fisher

04/11/2019

Reading in both the World Bank data and the CPIA data, that we want to merge into one

```
world_bank_indicators = read_csv("data/World Development Indicators.csv")
cpia_data = read_csv("data/CPIA Data.csv")
```

Cleaning the World Bank Data and turning it from a long format, to a wide format

```
world_bank_indicators = world_bank_indicators %>%
  select(
    Country = 'Country Name', Time, Series = 'Series Name', Value
  ) %>%
  drop_na(Country)

world_bank_indicators <- spread(world_bank_indicators, Series, Value)
```

Cleaning the CPIA Data and turning it from a long format, to a wide format

```
cpia_data = cpia_data %>%
  select(
    Country = 'Country Name', Time, Series = 'Series Name', Value
  ) %>%
  drop_na(Country)

cpia_data <- spread(cpia_data, Series, Value)
```

Joining the data together by Country and the Year

```
Clean_data = inner_join(cpia_data, world_bank_indicators, by = "Country", "Time")

Clean_data = Clean_data %>%
  select(-`<NA>.x`, -`<NA>.y`) %>%
  drop_na()
```

Selecting and renaming the variavles that we want to keep

```
Clean_data = Clean_data %>%
  mutate(
    Country = Clean_data$Country,
    Year = Clean_data$Time.x,
    Exports = as.numeric(
      Clean_data$`Exports of goods and services (constant 2010 US$)`
    ),
    FDI = as.numeric(
      Clean_data$`Foreign direct investment, net inflows (% of GDP)`
    ),
    GDP = as.numeric(Clean_data$`GDP per capita (constant 2010 US$)`),
    Capital = as.numeric(
      Clean_data$`Gross fixed capital formation (constant 2010 US$)`
    ),
    Savings = as.numeric(Clean_data$`Gross savings (% of GDP)`),
    Imports = as.numeric(
      Clean_data$`Imports of goods and services (constant 2010 US$)`
    ),
    Inflation = as.numeric(Clean_data$`Inflation, consumer prices (annual %)`),
    Aid = as.numeric(
      Clean_data$`Net ODA received (% of GNI)`
    ),
    Remitances = as.numeric(
      Clean_data$`Personal remittances, received (% of GDP)`
    ),
    Population_growth = as.numeric(
      Clean_data$`Population growth (annual %)`
    ),
    Population = as.numeric(
      Clean_data$`Population, total`
    ),
     CPIA_EM = as.numeric(
      Clean_data$`CPIA economic management cluster average (1=low to 6=high)`
    ),
     CPIA_SIE = as.numeric(
      Clean_data$`CPIA policies for social inclusion/equity cluster average (1=low to 6=high)`
    ),
     CPIA_PSMI = as.numeric(
      Clean_data$`CPIA public sector management and institutions cluster average (1=low to 6=high)`
    ),
     CPIA_SP = as.numeric(
      Clean_data$`CPIA structural policies cluster average (1=low to 6=high)`
    )
  )

Clean_data = Clean_data %>%
  select(
      Country,
      Year,
      Exports,
      FDI,
      GDP,
      Capital,
      Savings,
      Imports,
      Inflation,
      Aid,
      Remitances,
      Population,
      Population_growth,
      CPIA_EM,
      CPIA_SIE,
      CPIA_PSMI,
      CPIA_SP
    )

Clean_data = Clean_data %>%
  mutate(
    CPIA_Mean = (CPIA_EM + CPIA_SIE + CPIA_PSMI + CPIA_SP) / 4
  )
```

Creating the new variable of GDP growth and dropping NA's

```
Clean_data = Clean_data %>%
group_by(Country) %>%
mutate(Future_GDP = lead(GDP, order_by=Year)
)

Clean_data = Clean_data %>%
  mutate(GDP_growth = ((Future_GDP - GDP)/GDP)*100)

Clean_data = Clean_data %>%
  drop_na()
```

Saving the clean and combined data as an RDs and CSV

```
Clean_data %>%
  write_rds("data/clean_data.Rds") %>%
  write_csv("data/clean_data.csv")
```