Capstone R Work

Barrett Viator

2025-07-26

```
library(reticulate)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
```

```
# Portable setup for Python environment via reticulate
if (!requireNamespace("reticulate", quietly = TRUE)) {
  install.packages("reticulate")
}
library(reticulate)

env_name <- "retic_env2"

# Only create the environment if it doesn't exist
if (!env_name %in% virtualenv_list()) {
  message("Creating virtual environment '", env_name, "' and installing packages...")
  virtualenv_create(
    envname = env_name
  )

  # Install required packages after creation
  py_install(
    packages = c("pandas", "requests", "openpyxl", "pycountry", "wbgapi"),
    envname = env_name,
    pip = TRUE
  )
} else {
  message("Virtual environment '", env_name, "' already exists.")
}
```

```
## Virtual environment 'retic_env2' already exists.
```

```r
# Activate the environment
use_virtualenv(env_name, required = TRUE)

# Confirm setup
py_config()
```

```
## python:         /Users/raychandonnet/.virtualenvs/retic_env2/bin/python
## libpython:      /opt/homebrew/opt/python@3.13/Frameworks/Python.framework/Versions/3.13/lib/python3.
## pythonhome:     /Users/raychandonnet/.virtualenvs/retic_env2:/Users/raychandonnet/.virtualenvs/retic
## version:        3.13.5 (main, Jun 11 2025, 15:36:57) [Clang 17.0.0 (clang-1700.0.13.3)]
## numpy:          /Users/raychandonnet/.virtualenvs/retic_env2/lib/python3.13/site-packages/numpy
## numpy_version:  2.3.2
##
## NOTE: Python version was forced by use_python() function
```

```r
# Turn the python code into a string for Reticulate and import the python libraries needed - BJV
wb_undp_api_python_code <- "
import pandas as pd
import wbgapi as wb
import requests
import openpyxl
import pycountry

wb_series_codes = [
    'NY.GDP.PCAP.KD.ZG',
    'NY.GNP.PCAP.KD.ZG',
    'SE.XPD.TOTL.GD.ZS',
    'SE.ADT.LITR.ZS',
    'SE.COM.DURS',
    'SE.LPV.PRIM',
    'SE.LPV.PRIM.SD',
    'SE.SEC.CUAT.LO.ZS',
    'SE.SEC.CUAT.PO.ZS',
    'SE.SEC.CUAT.UP.ZS',
    'SE.TER.CUAT.BA.ZS',
    'SE.TER.CUAT.DO.ZS',
    'SE.TER.CUAT.MS.ZS',
    'SE.TER.CUAT.ST.ZS',
    'AG.PRD.FOOD.XD',
    'EN.POP.DNST',
    'EN.POP.SLUM.UR.ZS',
    'SP.RUR.TOTL.ZG',
    'EG.ELC.ACCS.ZS',
    'ER.H2O.FWST.ZS',
    'FX.OWN.TOTL.ZS',
    'SN.ITK.MSFI.ZS',
    'SH.XPD.CHEX.PC.CD',
    'SH.XPD.GHED.PC.CD',
    'SH.STA.WASH.P5',
    'SP.DYN.LE00.IN',
    'SH.UHC.SRVS.CV.XD',
    'IT.NET.BBND.P2',
    'IT.NET.USER.ZS',
    'GB.XPD.RSDV.GD.ZS',
```

```
    'IS.SHP.GCNW.XQ',
    'SI.POV.GINI',
    'VC.IHR.PSRC.P5',
    'CC.EST',
    'GE.EST',
    'PV.EST',
    'RL.EST',
    'VA.EST',
    'SM.POP.TOTL.ZS',
    'SI.POV.GINI'
]

# Create a data frame from the World Bank API - BJV

wb_df = wb.data.DataFrame(
    wb_series_codes,
    economy='all',
    time=range(1995, 2024),
    columns='series'
)

# Index needs to be reset in ordert for 'economy' and 'time' to be accessed and renamed - BJV
wb_df = wb_df.reset_index()

wb_df = wb_df.rename(columns={'economy': 'ISO3', 'time': 'year'})

# Remove YR from the 'year' column to leave just four digit years - BJV

wb_df['year'] = wb_df['year'].astype(str).str.replace('YR', '', regex=False)

# convert the year to a numeric from string for errors arising later - BJV
wb_df['year'] = pd.to_numeric(wb_df['year'], errors='coerce').astype('Int64')


# UNDP API Key and Request - BJV

undp_url = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe&y

response = requests.get(undp_url)

if response.status_code == 200:
    data = response.json()

df1 = pd.DataFrame(data)

# UNDP API Key and Request - BJV

undp_url2 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe&

response2 = requests.get(undp_url2)

if response2.status_code == 200:
    data2 = response2.json()
```

```
df2 = pd.DataFrame(data2)

# UNDP API Key and Request - BJV

undp_url3 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe

response3 = requests.get(undp_url3)

if response3.status_code == 200:
    data3 = response3.json()

df3 = pd.DataFrame(data3)

# UNDP API Key and Request - BJV

undp_url4 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe

response4 = requests.get(undp_url4)

if response4.status_code == 200:
    data4 = response4.json()

df4 = pd.DataFrame(data4)

# UNDP API Key and Request - BJV

undp_url5 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe

response5 = requests.get(undp_url5)

if response5.status_code == 200:
    data5 = response5.json()

df5 = pd.DataFrame(data5)

# UNDP API Key and Request - BJV

undp_url6 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe

response6 = requests.get(undp_url6)

if response6.status_code == 200:
    data6 = response6.json()

df6 = pd.DataFrame(data6)

# UNDP API Key and Request - BJV

undp_url7 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe

response7 = requests.get(undp_url7)

if response7.status_code == 200:
```

```
    data7 = response7.json()

df7 = pd.DataFrame(data7)

# UNDP API Key and Request - BJV

undp_url8 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe&

response8 = requests.get(undp_url8)

if response8.status_code == 200:
    data8 = response8.json()

df8 = pd.DataFrame(data8)

# UNDP API Key and Request - BJV

undp_url9 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubNe&

response9 = requests.get(undp_url9)

if response9.status_code == 200:
    data9 = response9.json()

df9 = pd.DataFrame(data9)

# UNDP API Key and Request - BJV

undp_url10 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response10 = requests.get(undp_url10)

if response10.status_code == 200:
    data10 = response10.json()

df10 = pd.DataFrame(data10)

# UNDP API Key and Request - BJV

undp_url11 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response11 = requests.get(undp_url11)

if response11.status_code == 200:
    data11 = response11.json()

df11 = pd.DataFrame(data11)

# UNDP API Key and Request - BJV

undp_url12 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response12 = requests.get(undp_url12)
```

```
if response12.status_code == 200:
    data12 = response12.json()

df12 = pd.DataFrame(data12)

# UNDP API Key and Request - BJV

undp_url13 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response13 = requests.get(undp_url13)

if response13.status_code == 200:
    data13 = response13.json()

df13 = pd.DataFrame(data13)

# UNDP API Key and Request - BJV

undp_url14 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response14 = requests.get(undp_url14)

if response14.status_code == 200:
    data14 = response14.json()

df14 = pd.DataFrame(data14)

# UNDP API Key and Request - BJV

undp_url15 = 'https://hdrdata.org/api/CompositeIndices/query?apikey=HDR-KOgLmhxaLjacLvGRi4oTPARBLx54ubN

response15 = requests.get(undp_url15)

if response15.status_code == 200:
    data15 = response15.json()

df15 = pd.DataFrame(data15)

all_dataframes = [df1, df2, df3, df4, df5, df6, df7, df8,
                  df9, df10, df11, df12, df13, df14, df15]

# Create a list of all data frames to be concatenated. -BJV
list_of_dfs = [df1, df2, df3, df4, df5, df6, df7, df8,
               df9, df10, df11, df12, df13, df14, df15]

# Concatenate all the data frames in the list. - BJV

combined_df = pd.concat(list_of_dfs, ignore_index=True)


# Create a df from a dictionary for saving file - BJV
undp_full_df = pd.DataFrame(combined_df)
```

```
# Country needed to split the ISO3 and the Country names - BJV
split_data = undp_full_df['country'].str.split(' - ', n=1, expand=True)

undp_full_df['ISO3 Code'] = split_data[0]
undp_full_df['Country Name'] = split_data[1]

# Drop the original 'Country' column - BJV
undp_full_df_split = undp_full_df.drop(columns=['country'])

# 'Value' was causing and error and needed to be converted to numeric - BJV
undp_full_df_split['value'] = pd.to_numeric(undp_full_df_split['value'], errors='coerce')

# Data needed to be pivoted for clarity when doing EDA - BJV

undp_full_df_pivoted = undp_full_df_split.pivot_table(index=['ISO3 Code', 'year'],
                           columns='indicator',
                           values='value')

undp_full_df_pivoted = undp_full_df_pivoted.rename(columns={'ISO3 Code': 'ISO3'})


# These are the columns that are relevant to our study from the total column options - BJV

columns_to_keep = [
    'hdi - Human Development Index (value)',
    'gii_rank - GII Rank',
    'rankdiff_hdi_phdi - Difference from HDI rank',
    'ihdi - Inequality-adjusted Human Development Index (value)',
    'coef_ineq - Coefficient of human inequality',
    'ineq_le - Inequality in life expectancy',
    'le - Life Expectancy at Birth (years)',
    'ineq_edu - Inequality in eduation',
    'ineq_inc - Inequality in income',
    'coef_ineq - Coefficient of human inequality'
]

undp_full_df_pivoted.index.rename(
    ['ISO3', 'year'],
    inplace=True
)

undp_df_filtered = undp_full_df_pivoted[columns_to_keep]

undp_df_filtered = undp_df_filtered.reset_index()

undp_df_filtered['year'] = pd.to_numeric(undp_df_filtered['year'], errors='coerce').astype('Int64')

# Joined Work Bank and UNDP Data Frames - BJV

undp_wb_filtered_full = pd.merge(undp_df_filtered, wb_df, on=['ISO3', 'year'], how='left')

# Use pycountry to map country names to ISO3 codes -BJV
iso3_to_name_map = {}
```

```
for country in pycountry.countries:
    iso3_to_name_map[country.alpha_3] = country.name

mapped_full_df = pd.DataFrame(undp_wb_filtered_full)

# Add the 'Country Name' column - BJV
mapped_full_df['Country Name'] = mapped_full_df['ISO3'].map(iso3_to_name_map)

cols = mapped_full_df.columns.tolist()

current_index_of_country_name = cols.index('Country Name')

# Pop country name - BJV

column_to_move = cols.pop(current_index_of_country_name)

# Move 'country name' to column 2
cols.insert(1, column_to_move)

mapped_full_df = mapped_full_df[cols]"


reticulate::py_run_string(wb_undp_api_python_code)

wb_undp_df_r <- py$mapped_full_df
```

```
# Example of the python code turning into a dataframe that works with tidyverse, etc. - BJV
glimpse(wb_undp_df_r)
```

```
## Rows: 5,974
## Columns: 54
## $ ISO3                                                          <chr> "AFG", "A~
## $ `Country Name`                                                <chr> "Afghanis~
## $ year                                                          <int> 1995, 199~
## $ `hdi - Human Development Index (value)`                       <dbl> 0.329, 0.~
## $ `gii_rank - GII Rank`                                         <dbl> NaN, NaN,~
## $ `rankdiff_hdi_phdi - Difference from HDI rank`                <dbl> NaN, NaN,~
## $ `ihdi - Inequality-adjusted Human Development Index (value)` <dbl> NaN, NaN,~
## $ `coef_ineq - Coefficient of human inequality`                 <dbl> NaN, NaN,~
## $ `coef_ineq - Coefficient of human inequality`                 <dbl> NaN, NaN,~
## $ `ineq_le - Inequality in life expectancy`                     <dbl> NaN, NaN,~
## $ `le - Life Expectancy at Birth (years)`                       <dbl> 52.103, 5~
## $ `ineq_edu - Inequality in eduation`                           <dbl> NaN, NaN,~
## $ `ineq_inc - Inequality in income`                             <dbl> NaN, NaN,~
## $ `coef_ineq - Coefficient of human inequality`                 <dbl> NaN, NaN,~
## $ `coef_ineq - Coefficient of human inequality`                 <dbl> NaN, NaN,~
## $ AG.PRD.FOOD.XD                                                <dbl> 67.87, 71~
## $ CC.EST                                                        <dbl> NaN, -1.2~
## $ EG.ELC.ACCS.ZS                                                <dbl> NaN, NaN,~
## $ EN.POP.DNST                                                   <dbl> 26.16536,~
## $ EN.POP.SLUM.UR.ZS                                             <dbl> NaN, NaN,~
## $ ER.H2O.FWST.ZS                                                <dbl> 59.04386,~
## $ FX.OWN.TOTL.ZS                                                <dbl> NaN, NaN,~
## $ GB.XPD.RSDV.GD.ZS                                             <dbl> NaN, NaN,~
```

```
## $ GE.EST                                                    <dbl> NaN, -2.1~
## $ IS.SHP.GCNW.XQ                                            <dbl> NaN, NaN,~
## $ IT.NET.BBND.P2                                            <dbl> NaN, NaN,~
## $ IT.NET.USER.ZS                                            <dbl> NaN, NaN,~
## $ NY.GDP.PCAP.KD.ZG                                         <dbl> NaN, NaN,~
## $ NY.GNP.PCAP.KD.ZG                                         <dbl> NaN, NaN,~
## $ PV.EST                                                    <dbl> NaN, -2.4~
## $ RL.EST                                                    <dbl> NaN, -1.7~
## $ SE.ADT.LITR.ZS                                            <dbl> NaN, NaN,~
## $ SE.COM.DURS                                               <dbl> NaN, NaN,~
## $ SE.LPV.PRIM                                               <dbl> NaN, NaN,~
## $ SE.LPV.PRIM.SD                                            <dbl> NaN, NaN,~
## $ SE.SEC.CUAT.LO.ZS                                         <dbl> NaN, NaN,~
## $ SE.SEC.CUAT.PO.ZS                                         <dbl> NaN, NaN,~
## $ SE.SEC.CUAT.UP.ZS                                         <dbl> NaN, NaN,~
## $ SE.TER.CUAT.BA.ZS                                         <dbl> NaN, NaN,~
## $ SE.TER.CUAT.DO.ZS                                         <dbl> NaN, NaN,~
## $ SE.TER.CUAT.MS.ZS                                         <dbl> NaN, NaN,~
## $ SE.TER.CUAT.ST.ZS                                         <dbl> NaN, NaN,~
## $ SE.XPD.TOTL.GD.ZS                                         <dbl> NaN, NaN,~
## $ SH.STA.WASH.P5                                            <dbl> NaN, NaN,~
## $ SH.UHC.SRVS.CV.XD                                         <dbl> NaN, NaN,~
## $ SH.XPD.CHEX.PC.CD                                         <dbl> NaN, NaN,~
## $ SH.XPD.GHED.PC.CD                                         <dbl> NaN, NaN,~
## $ SI.POV.GINI                                               <dbl> NaN, NaN,~
## $ SM.POP.TOTL.ZS                                            <dbl> 0.4264236~
## $ SN.ITK.MSFI.ZS                                            <dbl> NaN, NaN,~
## $ SP.DYN.LE00.IN                                            <dbl> 52.103, 5~
## $ SP.RUR.TOTL.ZG                                            <dbl> 4.7789179~
## $ VA.EST                                                    <dbl> NaN, -1.9~
## $ VC.IHR.PSRC.P5                                            <dbl> NaN, NaN,~
```

```
head(wb_undp_df_r)
```

```
##   ISO3 Country Name year hdi - Human Development Index (value)
## 1  AFG  Afghanistan 1995                                 0.329
## 2  AFG  Afghanistan 1996                                 0.334
## 3  AFG  Afghanistan 1997                                 0.338
## 4  AFG  Afghanistan 1998                                 0.338
## 5  AFG  Afghanistan 1999                                 0.347
## 6  AFG  Afghanistan 2000                                 0.351
##   gii_rank - GII Rank rankdiff_hdi_phdi - Difference from HDI rank
## 1                 NaN                                          NaN
## 2                 NaN                                          NaN
## 3                 NaN                                          NaN
## 4                 NaN                                          NaN
## 5                 NaN                                          NaN
## 6                 NaN                                          NaN
##   ihdi - Inequality-adjusted Human Development Index (value)
## 1                                                        NaN
## 2                                                        NaN
## 3                                                        NaN
## 4                                                        NaN
## 5                                                        NaN
## 6                                                        NaN
```

```
##   coef_ineq - Coefficient of human inequality
## 1                                        NaN
## 2                                        NaN
## 3                                        NaN
## 4                                        NaN
## 5                                        NaN
## 6                                        NaN
##   coef_ineq - Coefficient of human inequality
## 1                                        NaN
## 2                                        NaN
## 3                                        NaN
## 4                                        NaN
## 5                                        NaN
## 6                                        NaN
##   ineq_le - Inequality in life expectancy le - Life Expectancy at Birth (years)
## 1                                    NaN                                  52.103
## 2                                    NaN                                  52.830
## 3                                    NaN                                  53.212
## 4                                    NaN                                  52.487
## 5                                    NaN                                  54.532
## 6                                    NaN                                  55.005
##   ineq_edu - Inequality in eduation ineq_inc - Inequality in income
## 1                               NaN                             NaN
## 2                               NaN                             NaN
## 3                               NaN                             NaN
## 4                               NaN                             NaN
## 5                               NaN                             NaN
## 6                               NaN                             NaN
##   coef_ineq - Coefficient of human inequality
## 1                                        NaN
## 2                                        NaN
## 3                                        NaN
## 4                                        NaN
## 5                                        NaN
## 6                                        NaN
##   coef_ineq - Coefficient of human inequality AG.PRD.FOOD.XD    CC.EST
## 1                                        NaN          67.87       NaN
## 2                                        NaN          71.42 -1.291705
## 3                                        NaN          76.45       NaN
## 4                                        NaN          80.61 -1.176012
## 5                                        NaN          79.69       NaN
## 6                                        NaN          67.78 -1.271724
##   EG.ELC.ACCS.ZS EN.POP.DNST EN.POP.SLUM.UR.ZS ER.H2O.FWST.ZS FX.OWN.TOTL.ZS
## 1            NaN    26.16536               NaN       59.04386            NaN
## 2            NaN    27.23467               NaN       57.61255            NaN
## 3            NaN    28.29077               NaN       56.18125            NaN
## 4            NaN    29.37613               NaN       54.75702            NaN
## 5            NaN    30.49198               NaN       54.75702            NaN
## 6            4.4    30.86385               NaN       54.75702            NaN
##   GB.XPD.RSDV.GD.ZS    GE.EST IS.SHP.GCNW.XQ IT.NET.BBND.P2 IT.NET.USER.ZS
## 1               NaN       NaN            NaN            NaN            NaN
## 2               NaN -2.175167            NaN            NaN            NaN
## 3               NaN       NaN            NaN            NaN            NaN
## 4               NaN -2.102292            NaN            NaN            NaN
```

```
## 5               NaN       NaN            NaN            NaN            NaN
## 6               NaN -2.173946            NaN            NaN            NaN
##   NY.GDP.PCAP.KD.ZG NY.GNP.PCAP.KD.ZG    PV.EST    RL.EST SE.ADT.LITR.ZS
## 1               NaN               NaN       NaN       NaN            NaN
## 2               NaN               NaN -2.417310 -1.788075            NaN
## 3               NaN               NaN       NaN       NaN            NaN
## 4               NaN               NaN -2.427355 -1.734887            NaN
## 5               NaN               NaN       NaN       NaN            NaN
## 6               NaN               NaN -2.438969 -1.780661            NaN
##   SE.COM.DURS SE.LPV.PRIM SE.LPV.PRIM.SD SE.SEC.CUAT.LO.ZS SE.SEC.CUAT.PO.ZS
## 1         NaN         NaN            NaN               NaN               NaN
## 2         NaN         NaN            NaN               NaN               NaN
## 3         NaN         NaN            NaN               NaN               NaN
## 4           6         NaN            NaN               NaN               NaN
## 5           6         NaN            NaN               NaN               NaN
## 6           6         NaN            NaN               NaN               NaN
##   SE.SEC.CUAT.UP.ZS SE.TER.CUAT.BA.ZS SE.TER.CUAT.DO.ZS SE.TER.CUAT.MS.ZS
## 1               NaN               NaN               NaN               NaN
## 2               NaN               NaN               NaN               NaN
## 3               NaN               NaN               NaN               NaN
## 4               NaN               NaN               NaN               NaN
## 5               NaN               NaN               NaN               NaN
## 6               NaN               NaN               NaN               NaN
##   SE.TER.CUAT.ST.ZS SE.XPD.TOTL.GD.ZS SH.STA.WASH.P5 SH.UHC.SRVS.CV.XD
## 1               NaN               NaN            NaN               NaN
## 2               NaN               NaN            NaN               NaN
## 3               NaN               NaN            NaN               NaN
## 4               NaN               NaN            NaN               NaN
## 5               NaN               NaN            NaN               NaN
## 6               NaN               NaN            NaN                23
##   SH.XPD.CHEX.PC.CD SH.XPD.GHED.PC.CD SI.POV.GINI SM.POP.TOTL.ZS SN.ITK.MSFI.ZS
## 1               NaN               NaN         NaN      0.4264236            NaN
## 2               NaN               NaN         NaN            NaN            NaN
## 3               NaN               NaN         NaN            NaN            NaN
## 4               NaN               NaN         NaN            NaN            NaN
## 5               NaN               NaN         NaN            NaN            NaN
## 6               NaN               NaN         NaN      0.3853275            NaN
##   SP.DYN.LE00.IN SP.RUR.TOTL.ZG    VA.EST VC.IHR.PSRC.P5
## 1         52.103       4.778918       NaN            NaN
## 2         52.830       3.890502 -1.908540            NaN
## 3         53.212       3.688206       NaN            NaN
## 4         52.487       3.649521 -2.039301            NaN
## 5         54.532       3.611540       NaN            NaN
## 6         55.005       1.094174 -2.031417            NaN
```