HW 09: Python

David Gerard

2023-11-01

Instructions

- Write your solutions in this starter file. You should modify the "author" field in the YAML header.
- Only commit R Markdown and HTML files (no PDF files). Make sure you have knitted to HTML for your final submission.
- **Make sure to commit each time you answer a question.**
- Only include the necessary code, not any extraneous code, to answer the questions.
- Use only Python code (no R code).
- Use only Python functions from numpy, pandas, and seaborn (and functions that come with base Python).
- Do not change the path of any files.
- Learning objectives:
  - Numpy, Pandas, Seaborn

World Bank Data

The World Bank is an international organization that provides loans to countries with the goal of reducing poverty. The data frames in the dt folder were all taken from the public data repositories of the World Bank.

- country.csv: Contains information on the countries in the data set. The variables are:
  - `Country_Code`: A three-letter code for the country. Note that not all rows are countries; some are regions.
  - `Region`: The region of the country.
  - `IncomeGroup`: Either `"High income"`, `"Upper middle income"`, `"Lower middle income"`, or `"Low income"`.
  - `TableName`: The full name of the country.
- fertility.csv: Contains the fertility rate information for each country for each year. For the variables `1960` to `2017`, the values in the cells represent the fertility rate in total births per woman for the that year. Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
- life_exp.csv: Contains the life expectancy information for each country for each year. For the variables `1960` to `2017`, the values in the cells represent life expectancy at birth in years for the given year. Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
- population.csv: Contains the population information for each country. For the variables `1960` to `2017`, the values in the cells represent the total population in number of people for the given year. Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.

1. (1 pt) Use relative paths to load these data frames into Python.
2. (2 pts) These data are messy. The observational units in `fert`, `life`, and `pop` are locations in space-time (e.g. Aruba in 2017). Recall that tidy data should have one observational unit per row. Make these data tidy now.
3. (2 pts) Combine these data frames so that we have the fertility rate, population, life expectancy, and the region for each country in each year in a single data frame. The resulting data frame should look like this

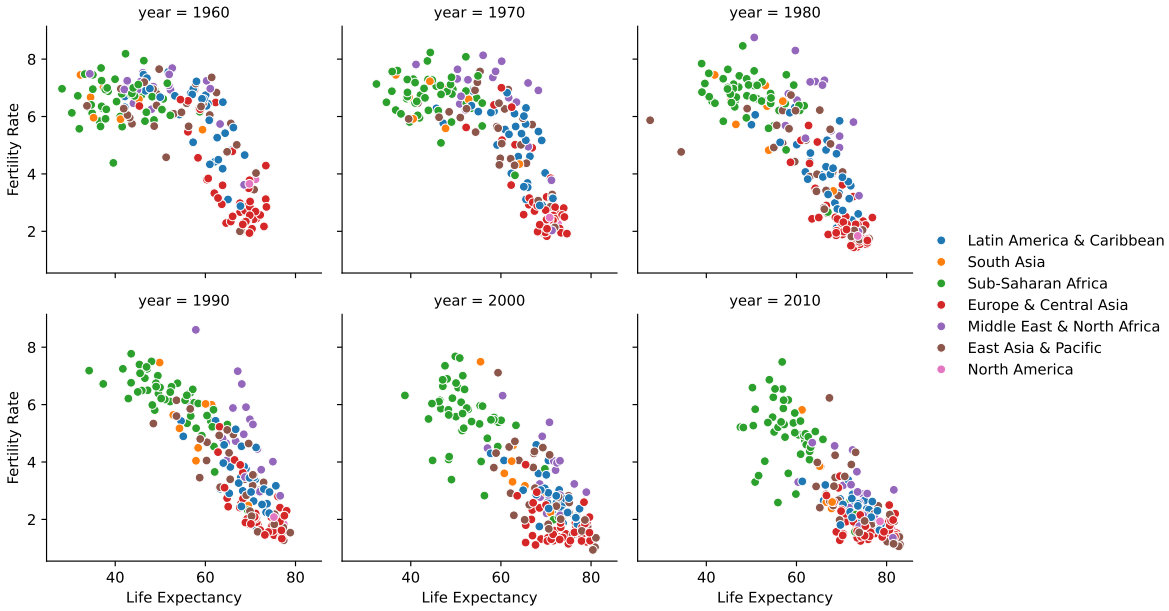
```
wb_df.head()
```

```
##   Country Name_x Country Code  ...  population                      Region
## 0          Aruba          ABW  ...     54211.0  Latin America & Caribbean
## 1    Afghanistan          AFG  ...   8996351.0                 South Asia
## 2         Angola          AGO  ...   5643182.0         Sub-Saharan Africa
## 3        Albania          ALB  ...   1608800.0      Europe & Central Asia
## 4        Andorra          AND  ...     13411.0      Europe & Central Asia
##
## [5 rows x 9 columns]
```

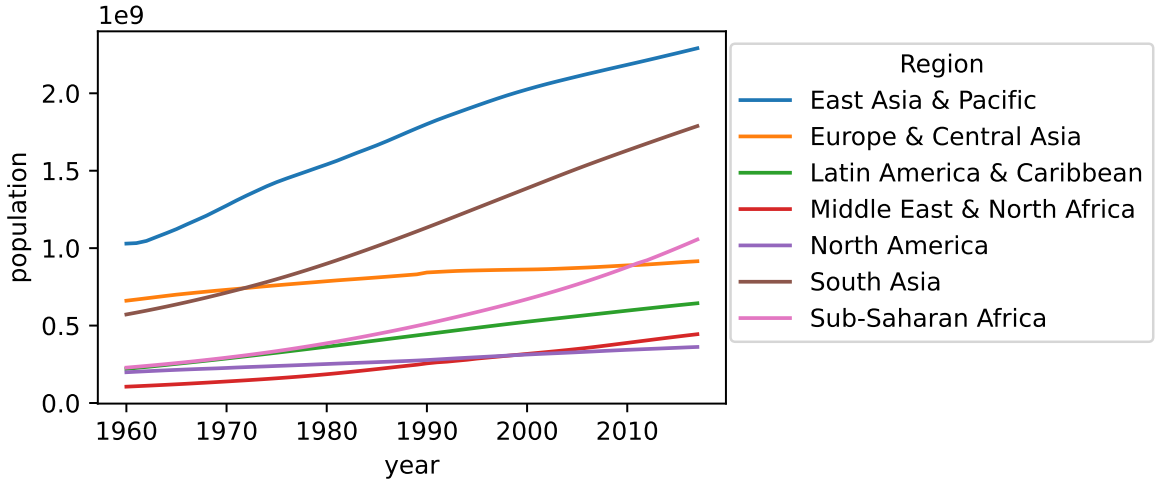
```
wb_df.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 15576 entries, 0 to 15575
## Data columns (total 9 columns):
##  #   Column          Non-Null Count  Dtype
## ---  ------          --------------  -----
##  0   Country Name_x  15576 non-null  object
##  1   Country Code    15576 non-null  object
##  2   year            15576 non-null  int64
##  3   fert_rate       14016 non-null  float64
##  4   Country Name_y  15576 non-null  object
##  5   life_exp        13997 non-null  float64
##  6   Country Name    15576 non-null  object
##  7   population      15147 non-null  float64
##  8   Region          12803 non-null  object
## dtypes: float64(3), int64(1), object(5)
## memory usage: 1.1+ MB
```

4. (2 pts) Make a scatterplot of fertility rate vs life expectancy, color-coding by region. Do this for the years 1960, 1970, 1980, 1990, 2000, and 2010. Facet by these years. Make your plot aesthetically pleasing. Your final plot should look like this:



5. (2 pts) Calculate the total population for each region for each year. Exclude 2018. Make a line plot of year versus total population, color-coding by region. Your final plot should look like this:



6. (2 pts) Make a bar plot of population vs region for the year 2017. Your final plot should look like this: