final-P3-sol

December 19, 2020

1 Academic Integrity Statement

As a matter of Departmental policy, **we are required to give you a 0** unless you **type your name** after the following statement:

> *I certify on my honor that I have neither given nor received any help, or used any non-permitted resources, while completing this evaluation.*

[TYPE YOUR NAME HERE]

2 Problem 3 (20 points)

Berkeley Earth is an independent nonprofit organization focused on using data science to understand contemporary environmental challenges. They have tabulated a monumental data set with detailed readings of the earth's surface temperature, with readings ranging back to the 1700s in many cases.

I have hosted a copy of one of their data sets at the following url:

`https://raw.githubusercontent.com/PhilChodrow/PIC16A/master/datasets/global_land_temperatures.`

The columns of the data include the month in which the temperature was read, the average temperature in that month, the uncertainty in that reading and the country in which the reading was recorded.

Create an interesting, professional plot using this data set, and write a paragraph describing your findings.

2.1 Requirements

- Your plot should make use of either **multiple axes** or **multiple colors** to distinguish between different parts of the data. You can of course use both, but this is not required to receive full credit.
- Your plot should include appropriate **labels** and **annotations**.
- It is permissible to use **for-loops** to loop over multiple axes or multiple groups within the data (such as countries), but not over individual entries of a data frame.
- Your explanation should include **a comment on NA values**, where/when they are most common, and how this could impact the interpretation of your plot.

- You are free to **supplement your main figure** with tables and additional figures. In this case, clearly label which figure is the main one.

2.2 Hints

- You will likely need to use`pandas` techniques to clean and simplify the data. `groupby` and `aggregate` are may be useful.
- You may need to figure out how to convert the `dt` column into a date that `pandas` can recognize.
- You may wish to extract information like the year from the `dt` column.
- I found the method `df.reset_index()` useful when working with the results of grouped aggregations.
- If you're not sure what to try, here's one possible suggestion. You are free (and encouraged!) to do something different, but a high-quality execution of the below suggestion is enough for full credit.
  - *Plot a trendline for each individual country in "the background" (e.g. light gray with transparency), and put the trendline for the overall average across countries in "the foreground" (e.g. black or red). Plot data before 1900 on one axis and data after 1900 on the second axis. Discuss the reasons for the rising mean in each of the two cases.*
- Other good questions you could try exploring:
  - In what countries are temperatures rising the most?
  - In what countries are we most confident about the temperature measurements in the 1800s? Are there any countries in which we have few or no measurements?

- Not sure how to transform a column in a particular way? Try to formulate your question as carefully as possible, and Google it!
- You might find the some of the cheatsheets to be helpful.

3 My Solution

[1]:
```
# import needed packages
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
```

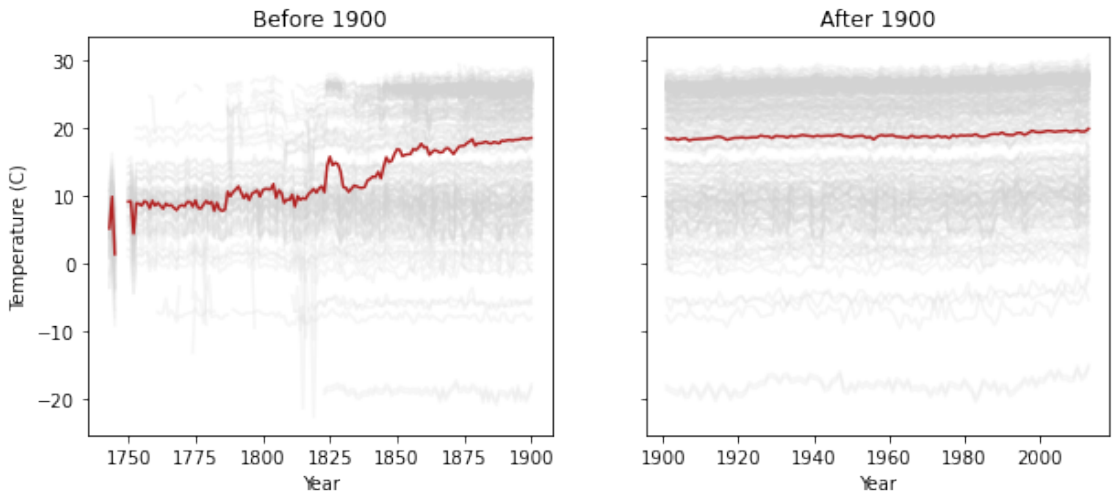
[2]:
```
# download the data
url = "https://raw.githubusercontent.com/PhilChodrow/PIC16A/master/datasets/
 ↪global_land_temperatures.csv"
df = pd.read_csv(url)
```

[3]:
```
# extract the year as a new column and
# add an indicator for whether the year is
# before or after 1900.
# note: extractin the year can also be done
# with string methods.
```

```
df['dt'] = pd.to_datetime(df['dt'])
df['year'] = pd.DatetimeIndex(df['dt']).year
df['divider'] = df['year'] > 1900
```

[4]:
```
# main plotting code

# create the figure
fig, ax = plt.subplots(1, 2, figsize = (10, 4), sharey = True)

def f(x, ax):
    """
    plot a single trendline of mean temperature per year vs. year,
    on the specified axis in light grey.
    Intended to work with one country's worth of data at a time.
    input x is a DataFrame, ax is a plotting axis.
    """
    ax.plot(x['year'], x['AverageTemperature'], color = "lightgrey", alpha = 0.
 ↪2)

def all_trendlines(df, ax):
    """
    plot (a) the mean temperature per year vs. year for each country in a light␣
 ↪grey
    and  (b) the mean across all countries in a highlighted red.
    df is an input DataFrame, ax is a plotting axis.
    """
    df.groupby(['Country', 'year']).mean().reset_index().groupby(['Country']).
 ↪apply(f, ax)
    ax.plot(df.groupby(['year'])['AverageTemperature'].mean(), color =␣
 ↪"firebrick")

# split the data into two pieces.
# could also use an apply-type solution here.
before = df[np.invert(df['divider'])]
after = df[df['divider']]

all_trendlines(before, ax[0])
all_trendlines(after,  ax[1])

# tidying up with labels
ax[0].set(ylabel = "Temperature (C)",
          xlabel = "Year",
         title = "Before 1900")

ax[1].set(xlabel = "Year",
          title = "After 1900")
```

[4]: `[Text(0.5, 0, 'Year'), Text(0.5, 1.0, 'After 1900')]`



What's especially interesting about these plots is that, from the trendline, it appears as though global temperatures rose a lot before 1900, and then only slightly after 1900. It's tempting to explain this through the Industrial Revolution, which took place in much of the world during this time and certainly increased global production of greenhouse gasses. However, this is a bit of an illusion—the amount of greenhouse gas produced in the entire 19th century is a drop in the bucket compared to production during the 20th and 21st.

This is where the missing data (NA) values come in. If you look at the lefthand plot, you can see that many countries, including a large number of very warm ones, only began to be measured in the 1800s. This leads the overall average to increase so much. By 1900, the data is quite consistent and so the trend is more steady (but still upward…)

Why do these warmer countries only begin to be measured in the 1800s? Well, that has to do with White colonialism. This was the era in which nations such as Great Britain, France, Germany, Italy, and the U.S. were racing to build empires in the Global South, especially in Africa and South Asia. The White colonists were often accompanied by scientists, who would record basic information such as temperature in those locations.

You weren't responsible for knowing the history, but (if you chose this route), I was looking for you to say something about how the missing data in warm countries made it difficult to interpret the sharp increase in the 19th century.