CSCI 140 Final Project: Endangered Languages

The data for this project is based on the data available at: https://www.theguardian.com/news/datablog/2011/apr/15/language-extinct-endangered. The data has been minorly adapted for this project, and is provided for you in the file GuardianLangs.csv. This data set contains information on endangered languages across the world, including the countries the languages are spoken in, the degree of endangerment, and the (estimated) number of speakers.

In addition to Guardianlangs.csv, there are two files included with this assignment: FinalLangs.py which contains a commented outline for the project with additional guidelines on the tasks you are required to complete, and countries.geo.json which is the .json file you will use for mapping (file source: https://github.com/python-visualization/folium/tree/master/examples/data). You must fill in your code to accomplish the tasks listed, generate the figures, and answer the questions in the assignment. **ANSWERS AND FIGURES WITH NO SUPPORTING CODE WILL RECEIVE NO CREDIT.**

Setup and Formatting (25%):

The data you will use is in the comma-separated value file GuardianLangs.csv. You will need to read in the data from the file into a pandas Data Frame. **NOTE: the file encoding is latin-1.** There are a few steps you need to take modify the Data Frame and the variables in it before completing analysis. Specifically:

1) Drop any columns where > 1/3 of the total observations are missing

2) Drop name in French, name in Spanish columns

3) Locate any rows missing a country code, use the coordinates to find the country (you do not have to do this with Python). Edit the observation to add country and country code for these rows.

QUESTION 1: For any rows for which you found missing a country code, list the language name, and the country and country code you filled in. Cite any sources you use. (2%)

4) Make ID a string, and make it the index

5) Drop any observation where Degree of Endangerment is equal to extinct or number of speakers is 0

6) Make Degree of Endangerment categorical, with ordered levels as from Vulnerable to Critically Endangered as indicated by the Guardian article

7) Rename columns: 'Country codes alpha 3' to 'Country_Code', 'Name in English' to 'Eng_Name', 'Degree of endangerment' to 'Degree_of_endangerment', 'Number of speakers' to 'Number_of_speakers'

8) Locate duplicate entries in the 'Eng_Name' column. Create a new data frame that contains only the observations with duplicated English names.

QUESTION 2: Choose ONE of these sets of duplicates and explain in 3-4 sentences the relationship (if any) between these languages and their names. Do you these these duplicates need to be removed based on what you found? Why or why not? Cite any sources you use. (2%)

Data Analysis (40%):

Here you will explore the relationship between the number of speakers and degree of endangerment. You will also modify the data frame to facilitate further analysis and mapping. The tasks to complete are:

1) Create a barplot showing the mean number of speakers grouped by degree of endangerment. **You must use a log scale for the numerical axis.**

2) Group the data by degree of endangerment and calculate and print descriptive statistics for the number of speakers for each category of degree of endangerment

QUESTION 3: What do you notice about the mean number of speakers over each category of degree of endangerment? Compare this with the range (the min to the max) for each category? Do the categories overlap? How does this correspond with the category definitions given in The Guardian article? Explain your answer in 3-5 sentences. (2%)

3) Create a **new** data frame that is a copy of the original, but only contains one country and country code per line. The new data frame should only contain the columns for the index (ID), English name, country, country code, and degree of endangerment. For example, suppose we consider a truncated version of the data frame that only contains 2 entries (NOTE: the index and columns in this example may be different from what you have after completing the QC steps):

1023	Sicilian	Italy	ITA	scn	Vulnerable	5000000.0	37.4399	14.5019
383	Low Saxon	Germany, Denmark, Netherlands, Poland, Russian...	DEU, DNK, NLD, POL, RUS	act, drt, frs, gos, nds, sdz, stl, twd, vel, wep	Vulnerable	4800000.0	53.4029	10.3601

Notice that the entry for Low Saxon has 5 countries listed. What we want to do is created 5 separate entries for Low Saxon such that each one lists only one country and country code, along with the English name and degree of endangerment. For example:

Dup0_1023	Sicilian	Italy	ITA	Vulnerable
Dup0_383	Low Saxon	Germany	DEU	Vulnerable
Dup1_383	Low Saxon	Denmark	DNK	Vulnerable
Dup2_383	Low Saxon	Netherlands	NLD	Vulnerable
Dup3_383	Low Saxon	Poland	POL	Vulnerable
Dup4_383	Low Saxon	Russian Federation	RUS	Vulnerable

Here, the indices were changed to give unique names for each duplicate – you do not have to have the same format for indices as shown here. The requirements are just that one country and one country code are listed per line, along with English name and degree of endangerment.

4) Make a table showing the number of languages by degree of endangerment by country code

Mapping (25%):

1) Create a new data frame (or Series) with country codes as the index and the count of languages (non-extinct) for that country in a column called 'Count'. You may have as many other columns as you like, you just need (at least) one column that contains the count.

2) Create a chloropleth map with each country colored in according to the count of languages. Your map should use the countries.geo.json file provided with this assignment. You should use a sequential palette (see http://colorbrewer2.org for palettes). Save this map as Final_Langs.html

QUESTION 4: You should notice that several countries on the map are nearly blank/barely colored in (you won't see this if you used a diverging or qualitative palette). Identify at least one of these that has languages corresponding to it in the data frame but is incorrectly colored in the map. If the number of endangered languages in the data frame is non-zero, why is the corresponding country not colored in? What could you do to fix this? Explain your answer in 3-5 sentences. (2%)

3) Make the correction you discussed in Question 4 for one (or more) of the impacted countries. Create and submit a new map that shows the correction with the filename Final_Langs_Corr.html.

QUESTION 5: Do you notice any geographic trends in the counts of endangered languages (you can use the map with blanks in it)? Which countries seem to have the most endangered languages? Do you think there are any factors not considered in the data set that could be influencing this? Explain your answer in 3-5 sentences. (2%)

What to submit:

-The filled-in FinalLangs.py file containing code to complete all of the tasks above
-The files containing your map: Final_Langs.html, Final_Langs_Corr.html
-A brief write-up containing the answers to questions in bold above and any comments you have on the project and any sources you used. This should be saved with the filename: YOURNAME.pdf

PROJECT CODE THAT DOES NOTE EXECUTE WITHOUT ERRORS AND/OR REQUIRES MANUAL INTERVENTION WILL BE SUBJECT TO A 10-15% PENALTY. TEST YOUR CODE!

Some helpful documentation links:

https://seaborn.pydata.org/ (Seaborn)
https://pandas.pydata.org/pandas-docs/stable/ (Pandas)
https://pandas.pydata.org/pandas-docs/stable/text.html (Working with text data in pandas)
https://pandas.pydata.org/pandas-docs/stable/groupby.html (Split-apply-combine, using groupby and agg)
https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.astype.html (Series astype)
http://folium.readthedocs.io/en/latest/quickstart.html(folium)