Assignment 2 (10 marks)

CIA Factbook Data Analysis and Visualization

I found an interesting dataset where interesting facts about the all the countries of the world have been posted by CIA. We have compiled it as (Country_data.csv), but still needs a fair amount of data cleaning.

https://www.cia.gov/library/publications/resources/the-world-factbook/docs/rankorderguide.html

Most of the fields are explained here (others are standard terms you can search for):

https://www.cia.gov/library/publications/resources/the-world-factbook/docs/notesanddefs.html

IMPORTANT NOTE:

- Please add explanations for each question/answer and your observations clearly in the Python notebook using the Headings and Markdown cells. If that is missing, it becomes very difficult for us to evaluate your solutions. Hence, solutions without Headings and clear explanations will not be graded.
- If any student is using unfair means (like copying the code of other student), they will get a zero for all assignments and D/F grade for the course. Please state your reference clearly at the end of your notebook.

QUESTIONS

Data Preparation (1 mark): Clean the dataset, i.e. convert the numerical values to numbers for the analysis. Please keep the rows and columns with "unknowns" as they have useful data in some of the columns which will be used in our analysis. Instead write your code such that it can handle the "unknowns" gracefully.

(1) (1 mark) Study the dataset carefully. Describe different aspects of this data via 5 most interesting plots. These must include:

- scatter plots
- histograms
- box and whisker type plot

For each plot, write **a paragraph in your notebook** showing the interesting stuff the visualization reveals. Please pay careful attention to the labels and titles of your plots.

Write at least 3 data modeling/analysis questions that come to your mind after looking at this data.

(2) (1 mark) Do a pairwise Pearson's correlation on all pairs of numerical variables to identify the top 20 pairs that correlate the strongest. Report your observation in the form of tables (Variable1, Variable2, Correlation_Coefficient). Repeat this with Spearman's rank correlation and compare the results. Can you explain the difference?

(3) (1 mark) Do permutation tests to determine the p-value of each observed correlation (on the top 100 list by Pearson's method) to test which are significant at a 5% level. What fraction of permutations produce at least this high a correlation?

Hint: https://piazza.com/class/kcyn9ji4405lo?cid=47 (There is a piazza post #47, on this topic)

(4) (2 mark) Analyze the distributions of different columns and select 3 candidate metrics (columns) which could be normally distributed. Now, perform Kolmogorov–Smirnov test to find which of these are close to normal (significance 5%).

Now identify 3 metrics (columns) which don't seem to be normally distributed. Perform the KS test again (significance 5%) and convince yourself that it actually works! For these 3 metrics, can you identify which known distribution it closely matches?

(5) (3 marks) Set up a regression model to predict the GDP - per capita (PPP) as a function of one or more of the other 'predictor' variables. You have several options to perform the feature selection. Use any 2 of the below:

a. Use OLS (statsmodel.api) and investigate the p-values, performing forward/backward/mixed selection as explained in the class to arrive at a good model.

b. Use the feature selection in Scikit learn library (implements a backward selection algorithm)

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

c. Use ElasticNet regression which is a form of **regularization** that forces some of the parameters 'β' to zero by adding terms in the Loss function. You can read about regularization from slides 52-58 of Lec9.pdf on Piazza. Usual linear regression tries to minimize MSE, but in order to select good features and reduce over-fitting, a penalty term is added to the objective function, so that useless features get dropped. We will study this in Lecture 10 on 9/7.

Using cross-validation, decide upon your final model. Assess the accuracy of your final model. Which countries are most above the forecast? Which are most below? Can you explain why?

(6) (1 mark) For apply methods like kNN (k-Nearest Neighbors), we have to find nearest neighbors. For this dataset, think of at least 2 methods of comparing countries for similarity or detecting nearest neighbors. In other words, design two similarity indices. For each of your index, print a table with the most similar and most different country for each country in the world. Do you find any interesting/unexpected pairs?

How to submit? This time, we just need you to submit your final notebook with answers to all the 5 questions in the assignment. Please remove any intermediate/irrelevant analysis from the final notebook submission.

Name of the Jupyter notebook should be: <INSTITUTEID>_ YourFirstName_Solution2.ipynb

Note: Students can feel free to reach out to the TAs and/or post questions on Piazza for clarifying any doubt regarding the assignment. If any student is using unfair means (like copying the code of other student), they will get a zero for all assignments and D/F grade for the course. Please state your reference clearly at the end of your notebook.