Instructor: Tales Imbiriba

Final Project

DS 5010
April 9, 2020

General guidelines:

Our final project aims at using the Python skills acquired in this class to analyse data, generate meaningful graphs and figures with the aim of extracting meaningful information and presenting it in a comprehensive and concise way. You should provide a source code and data used, and a written report.

Source Code: The source code should be well commented and organized, run without errors and generate all results presented in the written report.

Some instructions for writing Code:

1. For all major lines of code, please comment explaining what that line of code signifies. Writing comments along with your code is always a good practise
2. The code must be provided in '.py' format. If you are working with Jupyter Notebook, you can go to file-download-in py format and upload both ipynb and py files.
3. If you have plots, those plots would be shown in your final report. You can also submit the plots in JPG format along with code files.
4. The code should be executable. No indentation errors. The checker will execute the whole code and should not have any errors related to indentation or comments.
5. The code should also contain a README.md (or *.txt) file explaining what is each file and point out the main file that should be run.

Written report: The methodology adopted and the results obtained should be presented in a written report which must have the following sections:

1. Abstract: One **small** paragraph presenting the problem and its relevance, the methodology adopted and the results obtained.
2. Introduction: A brief presentation of the problem, objectives and organization of the rest of document.
3. Data Acquisition: present a description of the used dataset; describe the how the data was acquired and from where (with proper references), and describe any processing procedure used to prepare the data for analysis (e.g., cleaning procedure, imputation, etc).
4. Data analysis: Describe how did you extracted relevant information from the data sets (e.g., curve fitting, histograms, statistical analysis, other plots and figures, etc). In this section you should also present the results and discuss them.
5. Conclusions: present your final remarks.

Furthermore, the report should make comments on how important parts were implemented, and can present small parts of your python code. The report should be in PDF format.

Grading Policy

The project grade can can go from 0 to 100 divided in the following catogories:

Code (50 pts):

- (20pts) correctness;
- (20pts) generates and presents (graphs and prints) the results presented and discussed in the written manuscript;
- (10pts) organization and comments.

Report (50 pts):

- (25pts) Clarity and organization. Ideas should be clearly and succinctly presented. The manuscript should be organized following the suggested 5 organization points. The manuscript should be uniform regarding fonts and formatting. The ideas should be well organized and logically chained;
- (25pts) Results and discussion. The results should be meaningful and a discussion should be presented. The discussion should describe the results being presented (tables, figures, etc) and motivate conclusions and information that can be extracted from these results.

Proposed project: COVID-19 Data analysis

This project proposal consists in analyzing COVID-19 data provided by John Hopkins University and that can be found at https://github.com/CSSEGISandData/COVID-19. The goal of this project consists in

Our interest is to

1. Downloading data:

   https://raw.githubusercontent.com/datasets/covid-19/master/data/time-series-19-covid-combined.csv

2. Create a Data Frame from the CSV file.
3. Countries with multiple regions must be added to provide a single time-series for each country.
4. Select a subset of countries (e.g., 10 countries with highest number of confirmed covid-19 cases).
5. Analyze the data to answer the following questions:

   (a) Which countries present exponential growth and which countries already are already leaving exponential growth (for these countries Logistic function should present a better fitting). These can be done by fitting exponential and/or logistic functions or by graphical exploration using plots, histograms and other visualization tools.

   (b) Compare the statistics confirmed cases/deaths from the different countries and try to identify if there is any outliers. For comparing different countries the data should be somehow aligned by number of deaths or number of cases. Compute a mean behavior and dispersion with these aligned data. Make plots and/or scatter plots/box-plots, to analyse these behaviors. An interesting plot could be plotting the mean and $\pm 3\sigma$ (± 3 standard deviation) and plot the aligned evolution of each country. Feel free to explore here.

   (c) Provide a discussion of your findings and try to identify possible reasons for the discrepancy (if any) among the results obtained for each country.

   If you are adventurous you may try to perform some type of curve fitting and provide estimations for the near future. If that is the case you may try to fit exponential and/or logistic functions to the data.

   Exponential:

   $$y(t) = a * \exp(k * (t - t_0))$$

   where t is the time index t_0 is the time of the first confirmed occurrence, k is the growth constant and a is the initial number of cases at t_0.

   Logistic:

   $$y(t) = L/(1 + \exp(-k * (t - t_0))).$$

The data is organized in a CSV file with the following columns: Date, Country/Region, Province/State, Lat, Long, Confirmed, Recovered, Deaths. Note that some Countries do not have Province/Stats and have empty values for this field.