DTSC670: Foundations of Machine Learning Models
Assignment 2

Directions

The purpose of this assignment is to hone your pandas and Python skills. The truth is, data wrangling and preparation consumes the majority of time spent by a data scientist. Before predictions can be made, a model must be trained on clean data - that is, void of missing values, properly formatted, et cetera.

The following paper is included in Brightspace. This study aimed at determining the relationship of temperature to COVID-19 infection in the state capital cities of Brazil.

> David N. Prata, Waldecy Rodrigues, Paulo H. Bermejo. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. Science of The Total Environment. Volume 729. 2020. 138862. ISSN 0048-9697. https://doi.org/10.1016/j.scitotenv.2020.138862.

In this work, the authors considered a model similar to the following:

$$\log(y_{it}) = \theta_0 + \theta_1 x_{it}^3 - \theta_2 x_{it}^2 + \theta_3 x_{it} + \theta_4 T_i + \theta_5 d_i^2 + \theta_6 d_i + \theta_7 e_i + \varepsilon_{it}$$

where

Element	Description
y_{it}	The daily cumulative COVID-19 count in capital city i on day t
x_{it}	The number of days since the outbreak began in capital city i on day t
T_i	The average annual temperature for capital city i
d_i	The population density of city i
e_i	The population of city i
ε_{it}	Random Gaussian noise (Ignored when training a regression model)

It is not required to read the paper by Prata et al. to complete this assignment. However, if you have the time, then give it a glance - I find it interesting. In fact, I find the model used by these authors to be rather suspect. Here is one example underscoring my reservations: The authors claim that temperature significantly changes COVID-19 transmission, but the resolution of the temperature data that was considered by the model is far too coarse-grained. Is the *annual average temperature* of each capital city *really* sufficient for elucidating such a trend? Would this characterization change if the model considered the variation in temperature between months for each capital city?

The raw data I supply you with does include monthly average temperatures for these cities; if you want to try fitting a few models, indulge! However, you are not required to train any models

for this assignment. Note that I was unable to reproduce any of the trends published by the authors, but I employed different models than that used in the study.

Your task for this assignment is to perform the data preparation necessary to train the above model; however, you *will not* actually train any models for this assignment. You must construct a Jupyter notebook that reads in data from the supplied Excel file, then performs the necessary data manipulations to obtain two Pandas DataFrames.

The first DataFrame must be called `features`, which is your feature matrix. The `features` DataFrame must contain the following feature columns in this exact order: `days_cube`, `days_sq`, `days`, `temp`, `pop_dense_sq`, `pop_dense`, `pop`. These features correspond to the features present in the equation above.

- `days_cube` and `days_sq` are the cubed and squared values of `days`, respectively.
- `days` is the number of days since the outbreak began in the respective capital city, which always begins with zero and adds 1 day for each row of recorded data for each capital city. There are 27 capital cities in Brazil. Each capital city has 151 days of recorded data.
- `pop_dense_sq` is the squared value of `pop_dense`
- `pop` is the city population
- Population density is defined as the population of a city divided by the area in kilometers squared of the city.

The second DataFrame contains target values and must be called `response`, which are the accumulated case counts column for the capital cities. Your task is to construct these two Pandas DataFrames using the data provided.

You should begin by watching the video called Pandas DataFrame and completing the PandasDataFrame_template.ipynb. This video and notebook provide you with a review and some new materials about DataFrames - they will prepare you for completing Assignment 2.

Be sure to document your code! Include many comments indicating the semantics of your code! Ensure the names of your variables are meaningful.

Below is a simple rubric indicating how you will be graded:

Clarity and Comments 20%
How organized is your notebook? Do you have enough comments? Are the meanings of the computations clear?

Data Wrangling 80%
Were you able to successfully wrangle the COVID-19 data? Did you prepare the data for all 27 Brazilian state capitals for 151 days? Is your data correct and properly formatted?