CASE STUDY 031

[Python]

Global Temperature Change over Time – PART II

Difficulty Level: 3 of 3

NOAA stands for National Oceanic and Atmospheric Administration, formerly known as NCDC or National Climatic Data Center. NOAA is an agency of the United States government, under the Department of Commerce, that has the mission to understand and predict changes in climate, weather, oceans, and coasts, to share that knowledge and information about climate, and to conserve and manage coastal and marine ecosystems and resources.

The agency has records of temperature anomaly* since 1880 until today. The data used in this analysis has one data point to each month and its correspondent temperature anomaly. The complete data set can be downloaded at https://www.ncdc.noaa.gov/cag/time-series/global

** What is a temperature anomaly?*

The term temperature anomaly means a departure from a reference value or long-term average. A positive anomaly indicates that the observed temperature was warmer than the reference value, while a negative anomaly indicates that the observed temperature was cooler than the reference value.

More details can be found at https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php

Please refer to the Case Study 028, as this assignment is a continuation of it.

You are a data scientist working for a NGO that want to build a predictive model to forecast the temperature anomaly based on the observed data.

Your analysis should follow these steps:

1. Change the window of the rolling statistics to analyse the effect of this parameter in the output.

   a) Change it to a large value (120 months);

   b) Change to a small value (2 months);

   c) Change to 1 month;

   d) Explain the variation of the different versions of the plot;

   e) Explain why the first part of the lines representing the rolling mean and the rolling standard deviation aren't being shown;

2. The seasonal decompose that we plotted generates others series that you can work with, representing the components of your time series.

a) Assign to new variables each component trend, seasonal and residual;

b) Explore the new series.

c) Why there is null values in trends and residuals?

d) Remove the null values from the 3 lists and assure that they have the same shape (even the seasonal);

3. One of the key assumptions of the stationary is that there is no significant variance for each period.

   a) Plot a box plot for the data for each month for the residual component.

   b) Does the variance has a significant change over each bin?

4. Remove the seasonal component from the original series;

   a) Remove the same data points from the original series that we removed from the residual component in the step 2;

   b) Plot the new series;

5. Autocorrelation (ACF) shows the relationship with a data point with its previous values in time.

   a) Plot the ACF for the residual component + trend series (original - seasonal);

6. Partial autocorrelation (PACF) is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

   a) Plot the PACF for the residual component + trend series (original - seasonal);

7. Fit the model to data and predicts the next 6 months

   a) Create a new ARIMA model and fit the residual component + trend series to train the model;

   b) Predicts the interval between 01-01-2017 to 06-01-2017;

   c) Add the seasonal component to the predicted value;

d) Compare the results with the actual data from 2017;

e) Are the results of our predictions good?

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.