CLIMATE CHANGE: GLOBAL WARMING

There have been many studies documenting that the average global temperature has been increasing over the last century. The consequences of a continued rise in global temperature will be dire. Rising sea levels and an increased frequency of extreme weather events will affect billions of people. In this problem, we will attempt to study the relationship between average global temperature and several other factors. The file climate\_change.csv contains climate data from May 1983 to December 2008. The available variables include:  
  
a. Year: the observation year.  
b. Month: the observation month.  
c. Temp: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia ([http://www.cru.uea.ac.uk/cru/data/temperature/](https://www.google.com/url?q=http://www.cru.uea.ac.uk/cru/data/temperature/&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw2kYKOeP6ICdE5VmjpZBoSn)).  
d. CO2, N2O, CH4, CFC.11, CFC.12: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division ([http://www.esrl.noaa.gov/gmd/ccgg/dataproducts.html](https://www.google.com/url?q=http://www.esrl.noaa.gov/gmd/ccgg/dataproducts.html&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw0V3l5hz1y9iUnaM5vgk-Y-)).  
(i). CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)  
(ii). CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).  
e. Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA ([http://data.giss.nasa.gov/modelforce/strataer/](https://www.google.com/url?q=http://data.giss.nasa.gov/modelforce/strataer/&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw1YRiFOXfUp6YLB5HZjj9RL)).  
f. TSI: the total solar irradiance (TSI) in W/m (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARISHEPPA project website ([http://solarisheppa.geomar.de/solarisheppa/cmip5](https://www.google.com/url?q=http://solarisheppa.geomar.de/solarisheppa/cmip5&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw2ww18Ow6pbxMhMovLGBXSk)).  
g. MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation ([http://en.wikipedia.org/wiki/El\_nino](https://www.google.com/url?q=http://en.wikipedia.org/wiki/El_nino&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw3r-zp_ncN07sCpQggWmzfp)) (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division ([http://www.esrl.noaa.gov/psd/enso/mei/table.html](https://www.google.com/url?q=http://www.esrl.noaa.gov/psd/enso/mei/table.html&sa=D&source=editors&ust=1635060071625000&usg=AOvVaw3JXrZZZ3dpxRwCyR1zAHYx)).

Questions: -

1. Identify the dependent variable in the above data
2. Is this a Time-Series Data? Why or Why Not?
3. If you consider only the baseline what is the R2 of the model?
4. Run the raw regression and note whether the regression model is better than the baseline or not?
5. Identify the significant variables in the raw regression (one variable in each line)?
6. Which of the following assumptions are not fulfilled in the raw model?
   1. Normality
   2. Homoscedasticity
   3. Multicollinearity
   4. Absence of Correlated Errors
   5. Linearity
7. Which variables exhibit multicollinearity and why?
8. Which variables exhibit non-linearity and why?
9. Examine the residual plot and note your observations below
10. Modify the model and obtain your best model. What is its R2 and Adj R2?
11. Now Set seed by considering any number. Using the data mining approach obtain your best model and test it on testing data. Compare your models in terms of R2, Adj R2 and RMSE. Note the results below:
12. Compare the model obtained in Question 10 with that of Question 11 and note your observations below:
13. Write down your best model below:
14. If the residual plot shows autocorrelation then what steps can you take to overcome it?
15. Would you consider predicting using Time Series on this data? Why or Why Not?

Please provide the file to support the answers.

