Instructions:

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. This data comes from the website

https://crudata.uea.ac.uk/cru/data/temperature/

The file **climate_change.csv** contains climate data from May 1983 to December 2008. Read this data set into Python and answer the Questions in this Quiz. The variables in the data-set are-

--**Year:** the observation year.
--**Month:** the observation month.
--**Temp:** the difference in degrees Celsius between the average global temperature in that period and a reference value. This is the Dependent variable.
--**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.
--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)
--**CFC.11 and CFC.12** are expressed in ppbv (parts per billion by volume).
--**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.
--**TSI:** the total solar irradiance (TSI) in W/m2 (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.
--**MEI:** multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures).

Multiple choice (many answers)

1) We are interested in how changes in these variables affect future temperatures, as well as how well these variables explain temperature changes so far. To do this, first read the dataset climate_change.csv into Python. Then, split the data into a training set, consisting of all the observations up to and including 2006, and a testing set consisting of the remaining years. A training set refers to the data that will be used to build the model and a testing set refers to the data we will use to test our predictive ability.

   Next, build a linear regression model to predict the dependent variable Temp, using MEI, CO2, CH4, N2O, CFC.11, CFC.12, TSI, and Aerosols as independent variables (Year and Month should NOT be used in the model). Use the training set to build the model.

   Enter the model R2 (the "Multiple R-squared" value) on training set. (0.750)

2) Which variables are significant in the model? We will consider a variable significant only if the p-value is below 0.05. (Select all that apply.) MEI, CO2, CFC.11, CFC.12, TSI, Aerosols

3) Multiple choice (one answer)

Current scientific opinion is that nitrous oxide and CFC-11 are greenhouse gases: gases that are able to trap heat from the sun and contribute to the heating of the Earth. However, the regression coefficients of both the N2O and CFC-11 variables are negative, indicating that increasing atmospheric concentrations of either of these two compounds is associated with lower global temperatures.

1. There is not enough data, so the regression coefficients being estimated are not accurate.

Feedback: The linear correlation of N2O and CFC.11 with other variables in the data set is quite large. So there is Multicollinearity in the data. hence, the signs of some co-efficients are not making sense with intuition (or our general understanding). So, this is another way of catching Multicollinearity among IV. (Signs of IVs donot make intuitive sense)

2. Climate scientists are wrong that N2O and CFC-11 are greenhouse gases - this regression analysis constitutes part of a disproof.

3. All of the gas concentration variables reflect human development - N2O and CFC.11 are correlated with other variables in the data set.

All of the gas concentration variables reflect human development - N2O and CFC.11 are correlated with other variables in the data set.

4) Compute the correlations between all the variables in the training set. Which of the following independent variables is N2O highly correlated with (absolute correlation greater than 0.7)? Select all that apply.
CO2, CH4, CFC.12

5) Which of the following independent variables is CFC.11 highly correlated with? Select all that apply.
CH4, CFC12

6) Given that the correlations are so high, let us focus on the N2O variable and build a model with only MEI, TSI, Aerosols and N2O as independent variables. Remember to use the training set to build the model. Call this model2

Enter the coefficient of N2O in this reduced model: _____ (Enter in 2 decimals without rounding off)

Enter the model training R2: _____ (Enter in 2 decimals without rounding off)

0.02,0.72

7) True OR False.

We have observed that, for this problem, when we remove many variables the sign of N2O flips. The model has not lost a lot of explanatory power (the model R2 is 0.7261 compared to 0.7509 previously) despite removing many variables. As discussed in lecture, this type of behavior is typical when building a model where many of the independent variables are highly correlated with each other. In this particular problem many of the variables (CO2, CH4, N2O, CFC.11 and CFC.12) are highly correlated, since they are all driven by human industrial development. True

8) We have developed an understanding of how well we can fit a linear regression to the training data, but does the model quality hold when applied to unseen data?
Using the model (model2) produced earlier, calculate temperature predictions for the testing data set, using the predict function.

Enter the testing set R2: _____(Enter in 2 decimals without rounding off)
0.49