Exercise 1. ***Normal equation.***

To determine the set of parameters $\boldsymbol{\alpha}$ in linear regression (see lecture notes), we minimize the square Euclidean distance between model predictions $\hat{y}^{(j)} = {\mathbf{x}^{(j)}}^\top \boldsymbol{\alpha}$ and the actual values of the dependent variables $y^{(j)}$:

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^{m} \left[\hat{y}^{(j)} - y^{(j)}\right]^2 = \sum_{j=1}^{m} \left[{\mathbf{x}^{(j)}}^\top \boldsymbol{\alpha} - y^{(j)}\right]^2 = \|X\boldsymbol{\alpha} - \mathbf{y}\|_2^2 \,. \tag{1}$$

Show that the least-squares estimate $\boldsymbol{\alpha}^*$ of $\boldsymbol{\alpha}$ is given by the normal equation

$$\boldsymbol{\alpha}^* = (X^\top X)^{-1} X^\top \mathbf{y} \,. \tag{2}$$

Exercise 2. ***Climate change and global warming.***

The file `climate_change` (CSV) contains climate data from May 1983 to December 2008.[1] The available variables include:

- 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 data comes from the Climatic Research Unit at the University of East Anglia. 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.

  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. This data is from the Godard Institute for Space Studies at NASA.
- 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. This data is from the SOLARIS-HEPPA project website.

[1] The dataset was taken from https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/linear-regression/assignment-2/.

- 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). This data comes from the ESRL/NOAA Physical Sciences Division.

(a) Load the dataset into `Python`. 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.

(b) 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.

(c) Determine the coefficient of determination R^2 (see lecture notes) for the training data. How good are the temperature forecasts for the test dataset?