IC 272: Data Science III
Assignment 9: Autoregression

Consider the dataset provided in this assignment. This dataset describes the minimum daily temperatures over 10 years (1981-1990) in the city Melbourne, Australia. The units are in degrees Celsius and there are 3,650 daily observations. The source of the data is credited as the Australian Bureau of Meteorology and it can be downloaded from this link: https://bit.ly/2VvdAxG

Q1. Please answer the following questions:

(A). Please create a line plot of the dataset such that the X axis is the Date and the Y axis is the minimum daily temperature (in degrees Celsius). What do you observe in the line plot?
(B). Find the Pearson correlation coefficient between an observation and its previous (one timestep lagged) value. *Hint:* Create two columns one for the observations and another for the observations lagged by 1-day and then correlate the two columns (you may remove the first row where there would not be a lagged value for the first observation). Python function to use: *Corr()*

Q2. Following the procedure in Q1 (B) repeatedly, we could manually calculate the Pearson correlation values for each lag value of the observation and plot these correlations (Y axis) with lag (X axis) (this plot is called an autocorrelation plot). Fortunately, Python's statsmodels library provides a version of the plot_acf() function to make the same plot as a line plot. Use the plot_acf() function to make an autocorrelation plot up to a lag of 31 time periods. Please check whether the correlation value in Q1 (B) shows up for lag = 1 in the autocorrelation plot. What do you observe in the autocorrelation plot when the lag increases? *Hint:* Use the plot_acf(series, lags) function in Python.

Q3. Split the data into the following parts: training and test. The size of the training data is N-7 days and the size of the test data is the last 7 days (N are the total number minimum temperature observations in the dataset). Now, make a model that predicts data in the following manner: In the test data, any day's minimum temperature is equal to the last (just preceding) day's minimum temperature. Find the test RMSE for this model on this data. *Note:* Such a model is the simplest autoregression model and it is also called a *persistence model*.

Q4. An autoregression model is a linear regression model that uses lagged variables as input variables. For example, we can predict the minimum temperature value for the next time step (t+1) given the observations at the last 2 timesteps (t-1 and t-2). The autoregression model would look as follows: $X(t+1) = b_0 + b_1*X(t-1) + b_2*X(t-2)$, where b_0, b_1, and b_2 and autoregression coefficients.

Again, split the data into the following parts: training and test. The size of the training data is N-7 days and the size of the test data is the last 7 days (N are the total number minimum temperature observations in the dataset). Please use the AR class in the statsmodels library to develop an autoregression model on the dataset. The AR() class automatically selects an appropriate lag value using statistical tests, trains an autoregression model for the optimal

lag, and generates autoregression coefficients for the different lag terms in the model. Please report the optimal lag from the AR() class on training data (the model_fit.k_ar term) and the value of the autoregression coefficients for the optimal lag (the model_fit.params term). Now, use the developed autoregression model to generate the predictions on the last 7-days test data and compute the test RMSE. Please compare the test RMSE obtained with the one obtained in Q3.