Exercise 2 – Part B: Multiple Linear (Auto) Regression

For this exercise, we shall try to forecast store sales from the Rossman dataset. Particularly, you are required to build multiple linear regression models to forecast sales for the next 42 days. Please follow the following steps:

a. Find all the stores that have sales recorded for 942 days

b. Create a ***data*** matrix of the shape (#_of_stores, 942) for the daily sales record of these stores.

c. Use the first 800 stores in this ***data*** matrix for training and the rest for testing. Also split the sales data into 2 parts, the 1st part contains the information about the first 900 days of sales (these would be the features) and the 2nd contains the information about the last 42 days of sales (these would be the targets). The resultant matrices should be of the dimensions:

$$X_{train} = (800, 900)$$

$$X_{test} = (\#_{remaining\ stores}, 900)$$

$$Y_{train} = (800, 42)$$

$$Y_{test} = (\#_{remaining\ stores}, 42)$$

d. Iteratively build multiple linear regression models for column vectors of Y_{train}. You are allowed to use the numpy routines for calculating inverses, transposing of matrices and matrix multiplication. You would need to create 42 models in this case (1 model for each day in the target sales matrix)

e. Verify that you have learned $\beta_{0:900}$ for each of the 42 models and use these learned parameters to make predictions for each day ahead. In total 42 days.

f. Calculate and print the daily *RMSE* and *MAE* for all 42 sales values using test split (X_{test} as input). Also calculate and print overall average *RMSE* and *MAE*. (i.e. just the mean RMSE of all 42 models).

g. Use the following approaches and report the overall average *RMSE* and *MAE* for them:

   i. **Repeating last sale value per store:** Use the last recorded sales value of each store and repeat it for the next 42 days.

   ii. **Repeating mean value per store:** Repeat the mean of sales for the sales horizon.

   iii. **Repeating mean value per store per weekday:** For each of the 42 days ahead get their predictions as a mean of all sales recorded for that day of week in the past. For e.g., the prediction of Monday ahead should be the mean of all sales for this particular store on all previous Mondays.

h. Reason why or why not Linear Regression is a good choice for this task.