Module 5 Live Coding Assignment - Linear Regression: Interactions and Transformations

Throughout this assignment, please remember to import all the necessary libraries at the beginning of every exercise.

Question 1

Read the dataset “avocado.csv” and save it in a variable called “df”. Use the basic EDA functions to analyze de dataset.

Question 2

Drop erroneous columns from the dataframe. Use the function “to_datetime()” on the column “date” column to create the “day” and the “month” columns.

Question 3

Plot the Avocado?s “AveragePrice” using the “Date” column.
HINT:You will need to perform a groupby operation to do this.

Question 4

Get the correlation matrix and plot it for all the numerical columns in the dataframe.

Quesion 5

Write a function, “assert_normality” that uses the Shapiro-Wilks test to see if a column in our dataframe follows a normal distribution or not.

If not, use the Yeo-Johnson transformation to change the data so that it follows a normal distribution. Save your dataset as “df_norm”.

Question 6

From ‘scikit-learn‘, use RFE to get the best five variables to predict the AveragePrice for the avocados.

Question 7

Split your data into a training and a test dataset.

Use the best five variables from above perform a linear regression to predict the “AveragePrice” for the avocados. Get the MSE, MAE and RSME for your model.