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 the dataset.

Question 2

Drop unnecessary and erroneous columns from the dataframe. Do some feature engineering on the "date" column so you are able to use the day and the month columns in building our machine learning model later.

Question 3

Plot the Avocado's Average Price through the Date column. 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. Which variables are the most correlated with the "AveragePrice"? Which variables are highly correlated with others?

Quesion 5

Using the Shapiro-Wilks test build a function to see if a specific column follows a normal distribution or not. If not, using the Box-Cox or Yeo-Johnson transform them to follow a normal distribution. Save your dataset as "df_norm".

Question 6

Using the RFE method from “scikit-learn” get the best five (5) variables to predict the “AveragePrice” for the avocados.

Question 7

Divide your data into a training and a test dataset. Then using the best five (5) variables from above perform a linear regression to predict the “AveragePrice” for the avocados. Get the MSE, MAE y RSME for your model.