Mini_Project2_DAV_Question_rahul

The extract of the inferences from BigmartSales data is as follows.

1. Initially columns with no significant values like ID's where dropped using df.drop().
2. We used label encoder to 'Item_Fat_Content' column, and obtained 5 numeric equivalents
3. Similarly for columns "Item_Type","Outlet_Type","Outlet_Location_Type","Outlet_Type" we used Ordinal encoder
4. Null value treatment
    a. Outlet_Size was imputed with mode() value
    b. Item_Weight was also imputed with mode() value
5. df.boxplot() drew box plot for all columns but due to data being of different ranges box plot is not clear
6. For this raw df we performed training and testing with 80-20 ratio using Linear Regression model
    a. Root Mean Squared Error (RMSE): 1192.529066514257
7. Apply StandardScaller and split data set.
    a. Interesting now box plot has visibility as all the data has been scaled to standardScaler
8. Now if we applied Linear Regression fit and checked for RMSE its reduced to value under 1
    a. Root Mean Squared Error (RMSE): 0.699341128456553
9. Similarly
    a. Linear Regression (RMSE): 0.7124908079587138
    b. MinMaxScaler (RMSE): 0.09228729251525904
    c. RobustScaler (RMSE): 0.5284709217082865
    d. MaxAbsScaler (RMSE): 0.09316878002641989
    e. Normalizer (RMSE): 0.07720579872579159
10. Post scaling RMSE has reduced but that doesn't mean model is working better it just scaled down. Usually scaling doesn't have impact on linear regression.
11. Finally a box plot with legends is drawn to show the differences RMSE across standardization.