PSL Project2:

This project aims to construct a predictive model capable of forecasting future weekly sales for individual departments within 45 Walmart stores. The model will leverage historical weekly sales data to make accurate predictions.

Technical Details:

1. **Data Pre-processing**

   As shown by the weekly sales data, the trends for a department across different stores appear consistent. In order to diminish noise, Singular Value Decomposition (SVD) was employed to pre-process train data.

   To further prepare the training and testing datasets for modeling, the "Date" column was decomposed into two new features: "Yr" and "Wk". The "Yr" feature, representing the year, was treated as a numerical variable. The "Wk" feature, representing the week of the year, was categorized into 52 levels.

   Some levels in the training set do not have observations. When dealing with a non-full rank training matrix, a R-based approach was employed. Basically, starting with the last column, I assessed if it can be expressed as a linear combination of preceding columns using the least squares solution. If the residuals are almost zero which means the column can be represented as a linear combination of previous columns, that column is removed. I used the same process to work backwards all the way to the intercept column to refine the training matrix. The matching columns are then removed from the testing matrix.

   Other missing values found during the implementation were filled in zeros.

2. **implementation**

   For each unique combination of department and store, a separate linear regression model was trained using the formula Y ~ Yr + Yr_squared + Wk. This model was then used to predict future weekly sales for that specific department-store combination.

   To boost the running time, an efficient method was employed for implementation. Basically, a unified design matrix was constructed for the entire dataset using Patsy's dmatrix function. This matrix was subsequently partitioned into subsets, each corresponding to a specific department-store combination. This approach eliminated the need for repetitive data subsetting and design matrix creation during the iterative model training process. The predicted sales values for each department-store combination were concatenated, and then merged with the original test data.

Performance Metrics:

The computer system used is: MacBook Pro 13.3" Laptop - Apple M2 chip - 24GB Memory - 1TB SSD (Latest Model) - Silver.

The performance and the execution time on the 10 splits are as follows:

Folder No.	**WMAE** **(Weighted Mean Absolute Error)**	**Execution Time**
1	1958.221	24.97856903076172
2	1373.363	25.576316833496094
3	1388.133	27.48302698135376
4	1537.629	28.52991819381714
5	2330.364	30.358338117599487
6	1644.030	31.261943101882935
7	1620.090	31.415032863616943
8	1360.600	32.764655113220215
9	1343.752	34.463091135025024
10	1341.333	34.43692994117737
Average	**1589.751**	