# PDP – Assignment 2 (#9 – Walmart Analysis)

Name: Mohammed Al-selmi

Student number: 669668

Github URL for the assignment: https://github.com/mm55355/Hadoop-Assignments

## Steps:

1. Before starting with assignment, I downloaded the Walmart dataset from Kaggle.
2. After that, I had to import the necessary libraries such as panda, sklearn...etc. Then read the dataset and explore it for instance, check for missing values.
3. Before making a decision the predictive features for the model, I made sure to pre-process the dataset. Additionally, I visualize the relationship between features and the target to determine the top three predictive features.
4. After determining the features, I created the pipeline with the model, then I evaluated the model.

## Explanation of choices made:

Regarding the choices made in feature design, pre-processing, and model selection, here are some explanations:

* **Feature Design**: The ‘Date’ column is transformed into separate ‘Year’, ‘Month’, and ‘Day’ features to capture any patterns in the data. This allows the model to potentially learn different behaviours based on these time-related features. Furthermore, in order to display the heatmap it requires date to be datetime instead of object which was the original type of the ‘Date’ column.
* **Pre-processing:** The numeric features are selected and pre-processed using a transformer pipeline. Missing values are imputed using the median strategy, and the data is scaled using StandardScaler. This pre-processing ensures that all numeric features are on a similar scale and handles missing values appropriately.
* **Model Selection:** Linear Regression is chosen as the model for this task. This choice might be influenced by various factors, such as the assumption that there might be a linear relationship between the features and the target variable.

## Assignment discussion:

To discuss the assignment requirements, let's break it down into three main parts:

1. **Pre-processing necessary features**: In the code, pre-processing steps are applied to the dataset. The ‘Date’ column is converted to a datetime object, and additional features such as ‘Year’, ‘Month’, and ‘Day’ are extracted from it. This pre-processing step allows the model to potentially capture any patterns or seasonality in the data.
2. **Predicting next week's sales and finding top 3 predictive features:** The assignment requires predicting next week's sales. In the code, a Linear Regression model is trained using the pre-processed features and the ‘Weekly\_Sales’ target variable. The model is then used to predict the sales for a specific row in the testing data. To find the top 3 predictive features, three different methods of measuring predictiveness can be employed. I used four different methods to measure the correlation between the numerical features and target column. The methods used were the following:
   1. Heatmap
   2. Scatter plots for each feature
   3. Feature importances
   4. SelectKBest
3. **Reporting score/accuracy in at least 2 different formats:** The code calculates and prints the model score using the clf.score() function, which represents the coefficient of determination (R^2) of the Linear Regression model. The R^2 value indicates the proportion of the variance in the target variable that is predictable from the features. Besides that, different methods were used such as, Mean Absolute Error Value, Mean Squared Error Value, Median Absolute Error Value, and Adjusted Determination Coefficient.

## Pipeline runtime:

A screenshot of a computer

Description automatically generated with medium confidence