Group Name: Team Rocket
Project Name: Retail Forecasting
Member: Jie Heng Yu
Email: jiehengyu09557@gmail.com
Company: DataGlacier
Batch Code: LISUM26
Specialization: Data Science

Problem: X is a company that has a beverages business in Australia. They sell their products through various super-markets & also engage in heavy promotions throughout the year. Their demand is influenced by various factors like holidays & seasonality. They need a forecast for each of their products at the item level in weekly buckets.

Business Understanding: The objective is to build a multivariate machine learning model that will be able to forecast sales weekly. The data that is required to build the model may need to be re-coded so that the dates are in weekly buckets.

Project LifeCycle:
Understanding of the problem -> Data understanding -> Cleaning the data -> Exploratory data analysis -> EDA presentation/model selection -> Building the model -> Model presentation
< Deadlines >

- November 19, 2023
    - This pdf, Data intake report, GitHub repository link
- November 26, 2023
    - PDF of data understanding, explaining the following: Type of data, Problems in data, Approach to overcome problems in data, GitHub repository link
- December 2, 2023
    - At least two techniques to clean the data saved on a ipynb file, GitHub repository link
- December 9, 2023
    - Exploratory data analysis, GitHub repository link
- December 16, 2023
    - EDA presentation, model recommendation, GitHub repository link
- December 23, 2023
    - Model benchmarking, GitHub repository link
- December 30, 2023
    - Model presentation, GitHub repository link

Data Understanding

1. *What type of data have you got for analysis?*
   The data that I am working with is numeric, with the exception of the features: Product, date, & Price.Discount. The exceptions are strings. Features In.Store.Promo, Catalogue.Promo, Store.End.Promo, Google_Mobility, Covid_Flag, V_DAY, EASTER, & CHRISTMAS are boolean values: 1 denoting True, 0 denoting False.

2. *What are the problems in the data (number of NA values, outliers, skewed, etc)?*
   The problems I see in the dataset are the features Price.Discount…. & date. The Price.Discount…. feature needs to be renamed. The date feature needs to be re-coded into weekly buckets.
3. *What approaches will you apply to your dataset to overcome problems like NA values, outliers, etc. & why?*
   If there are NA values, I will remove the rows with them. It's better to train a model when we don't have missing data. Although there are many methods to resolve missing data, I feel that doing so will not reflect the true population values of the features. I believe that for this project, the model could be slightly overfit, since only this company will be using this data for its forecasting. Outliers in the sales feature could be the results of promotions, seasonality, holidays, etc. & I want to take that into account when training the model. As such, I will try a model with outliers & without outliers. When building the model without outliers, only extreme outliers will be removed. I believe the systematic way to identify extreme outliers in a statistical research setting is anything over 75th quartile + 3 IQR or anything under 25th quartile - 3 IQR. Anything within will be retained. Extreme outliers will affect the model training process. This model that has removed outliers will be more generalizable. We will determine which is better with the models' predictive accuracy & context.