**Job-a-Thon September**

**Approach File**

1. **A brief on the approach, which you have used to solve the problem.**

My approach was very simple. First I did some basic data exploration, then manually converted categorical columns in to numerical for better calculations, then created a new column and signed value from 0-4, this will divide data into 5 equal parts. This can work as another way of applying K-fold cross validation.

Then selected useful columns- dropping ‘ID’ column for very high variability, dropped ‘#Order’ column as in wasn’t used in Test dataset, dropped ‘Kfold’, ’Date’ column as they are not of much help, then finally dropped ‘Sales’ column as it is our target column so we will access it later.

Then performed One-hot encoding on each fold and trained model by using XGB Regressor algorithm. I have separately used optuna – a hyper parameter tuning framework to figure out best parameters for this model. Then finally made my final prediction using this model.

1. **What data-preprocessing / feature engineering ideas really worked? How did you discover them?**

Feature Engineering : By manually exploring the dataset, I have found some columns like 'Store\_Type', 'Location\_Type', 'Region\_Code' and 'Discount' which are categorical type, so I replaced their values with appropriate numbers.

Then I create another column ‘kfold’ and assigned values 0-4, dividing dataset into 5 equal parts, So that I can use it for cross validation later.

Data-preprocessing : I have used One-hot encoding for final preprocessing.

1. **What does your final model look like? How did you reach it?**

My final model is made using XGB Regressor algorithm. I have tried many algorithms like Random forest algorithm and Gradient boosting algorithm before reaching it. For final modeling I have separately use Optuna framework to find best parameters for this model then used them for final modeling.