Retail Forecasting for a Large Beverage Company in Australia

Problem Statement:

A large beverage company in Australia, which sells products through various supermarkets and runs frequent promotions, requires weekly demand forecasts at the item level. Their current in-house software struggles to accurately forecast demand due to complex patterns in the time series data, which are influenced by trends, seasonality, and holidays. As a result, the company seeks to explore AI/ML-based forecasting solutions to replace their existing system.

Goal:

To develop accurate and efficient AI/ML-based multivariate forecasting models that can replace the company's in-house software, improving demand forecasting by handling trends, seasonality, and other influencing factors. The models should demonstrate best-in-class accuracy and provide explainability regarding the impact of different variables.

Approach:

1. **Data Cleaning:**
    - **Checking for missing values:**

      For every product except product 6, all have some data entry with 0 sales in the dataset. So remove those rows with 0 sales. Also if we look at SKU6 alone, it has no missing data, just an earlier end date.

    - **Addressed outliers:**

      For numerical variables, we can look for outliers and skewness. We can see that for all products, Sales is right skewed and Google_Mobility is left skewed, while Price Discount (%) is more randomly scattered. The majority of potential outliers are large spikes in sales. These spikes could be caused by factors like promotions, holidays, or other events leading to increased sales. They also display a seasonal trend. The Z-Score and IQR methods are both detecting these positive sales spikes as outliers.These spikes are not "errors" in the data instead represent meaningful events (e.g., successful promotions, holiday seasons). So they are important data point for us. We can't see any spikes which may not cause by any holidays or promotion result in any potential outliers. So we can't consider those skewness as outliers or data error.
      To determine why "Google_Mobility" is left-skewed and whether it's influenced

by factors like COVID, holidays, or promotions. The drop of Google_Mobility is primarily driven by the COVID-19 pandemic period. A strong negative correlation with Google_Mobility. During these times, mobility was heavily restricted, which explains the drop and the skewness. So we can't cosider those extremely low data points as outliers. We will have to keep those for our further predictions and trends.

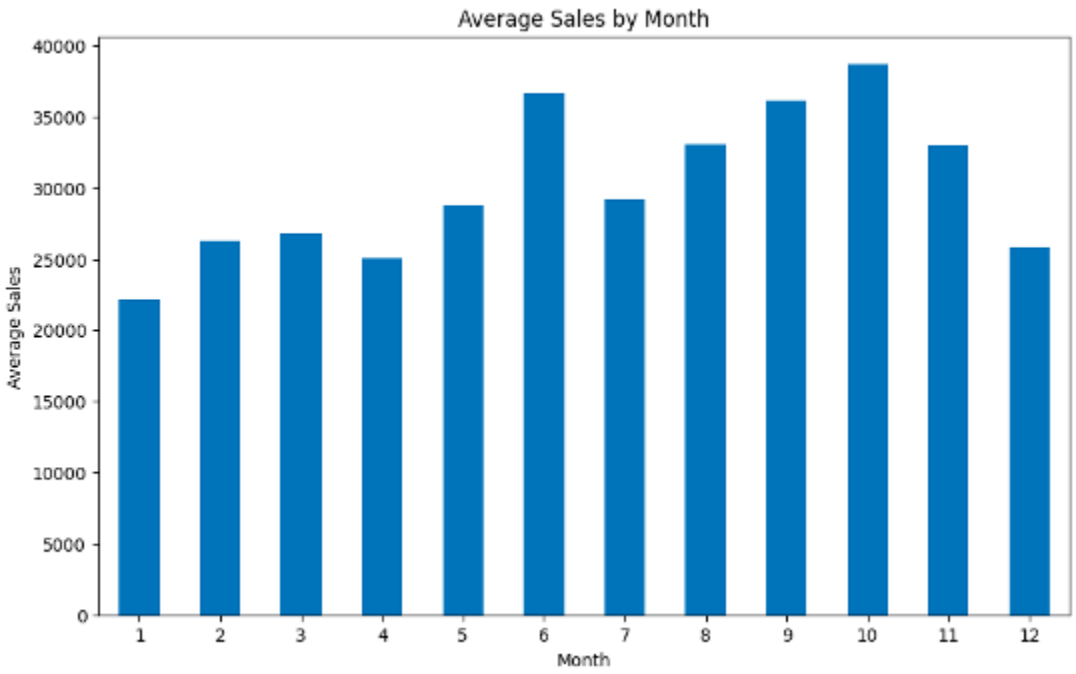
2. **Outlier Handling:**

As we cannot remove outliers, except some zero values at the end. We have to handle the rest of the outliers using different outlier handling techniques. We tried transformation techniques such as log, square root, box-cox, and Yeo-Johnson transformation to reduce the skewness as we can't remove those data points. But even after transformation, the data still skewed a lot.

3. **EDA :**

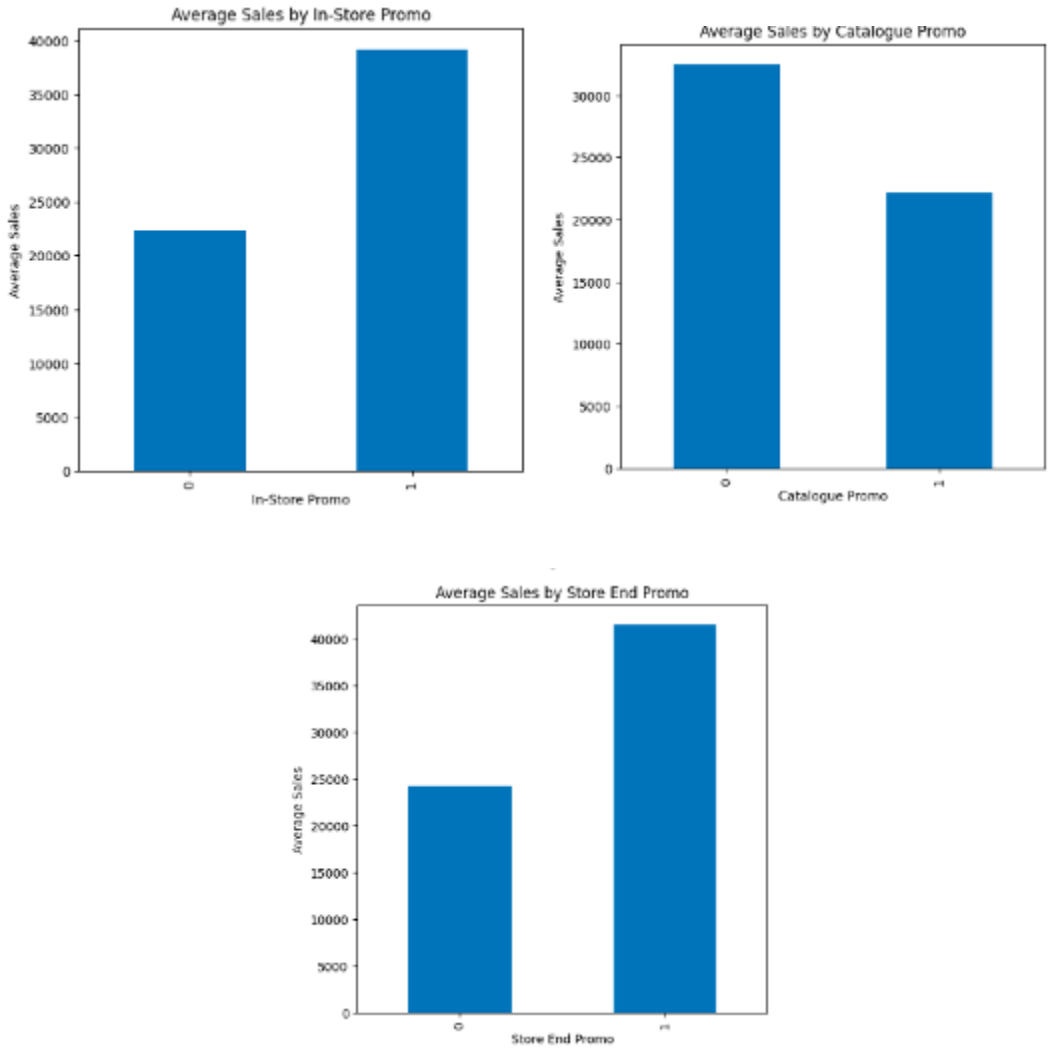
General Analysis for all products

1. What is the seasonal pattern of average monthly sales, and how does seasonality influence product demand?



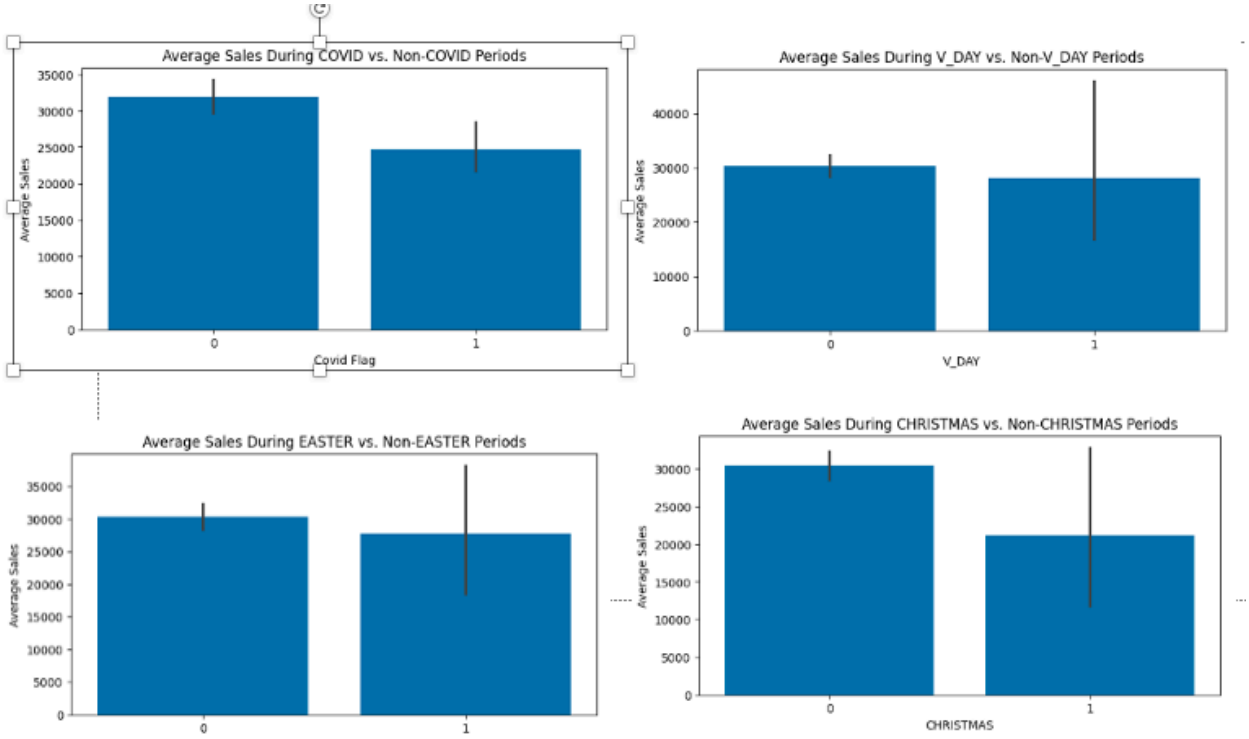
Sales are **highest in June (Month 6)** and **October (Month 10)**, indicating strong demand during mid-year and pre-holiday periods. The sales peaks in June and October may be driven by specific holiday seasons or promotional campaigns. There is a noticeable dip in **December (Month 12)** despite the holiday season, which could suggest the need for more focused promotions or inventory adjustments during this time.

2. How do different promotion strategies (In-Store, Catalogue, and Store End) influence sales performance?



In-store and store end promotions are highly effective strategies for boosting sales, while catalogue promotions seem to have a more neutral or even negative impact on sales performance. The company should focus more on physical, in-store promotional strategies to drive demand.

3. What is the impact of covid/non-covid time on sales and how do sales trends differ during holiday periods compared to non-holiday periods on sales ?



COVID seems to have a negative impact on sales, likely due to restrictions, changes in consumer behavior, or disruptions in supply chains.

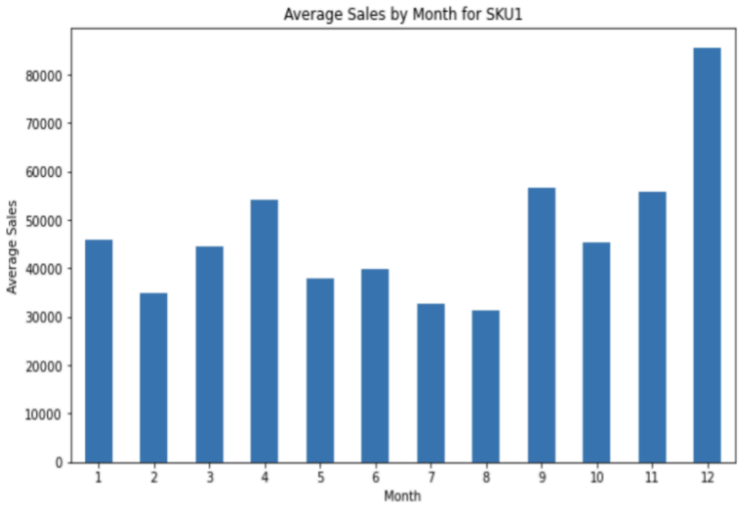
Overall, holiday periods show no significant positive impact on sales compared to non-holiday periods.

The negative trend during Christmas could suggest that customers may purchase well in advance or opt for alternatives.

Individual Analysis for all products:

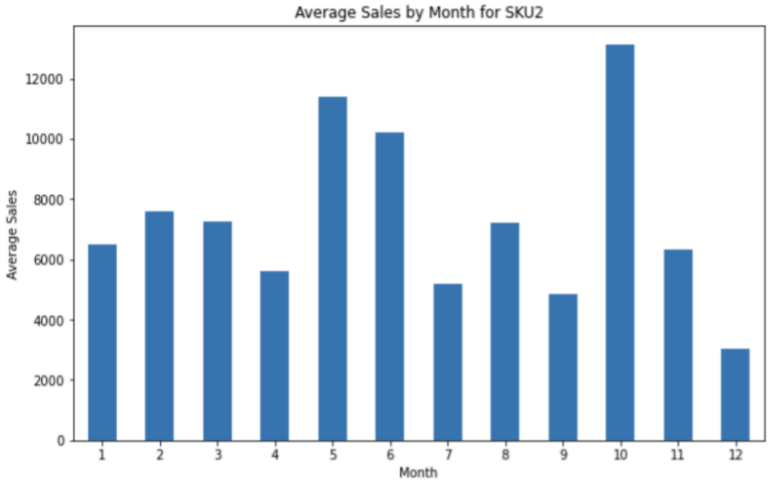
Average monthly sales per product

SKU1 Monthly Average Sales



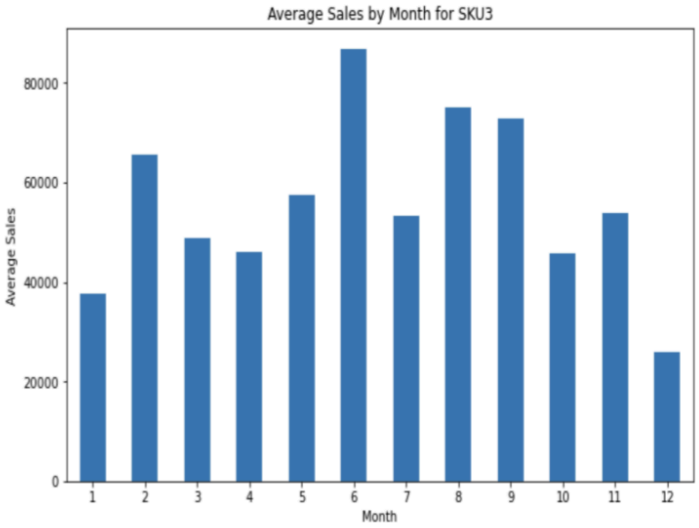
The Average sales of SKU1 fluctuates throughout the year, in november and december the sales are high; this could be influenced by high demand or other factors like promotions or the festive season.

SKU2 Monthly Average Sales



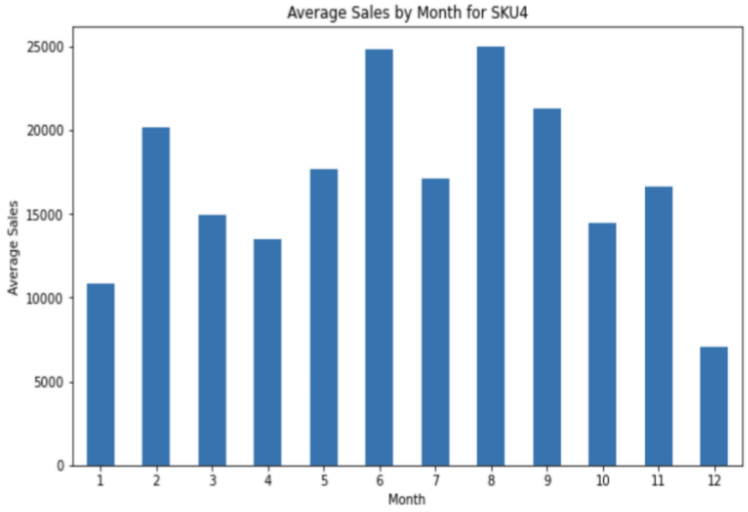
The average sales of Sku2 are high in october, may and june indicating the high product demand in those months and the lowest in december.

SKU3 Monthly Average Sales



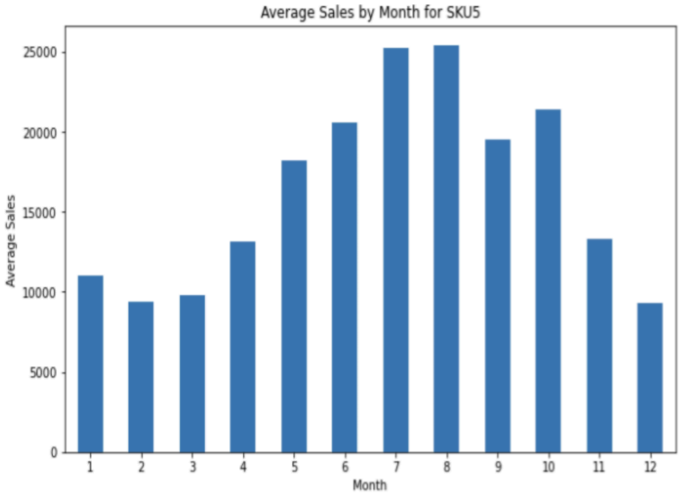
Sku3 sales well throughout the year, some months performing well than the others; for instance june, august, september and february the product's demand is high and quite low in december. This demand could be influenced by some factors like seasonality.

SKU4 Monthly Average Sales



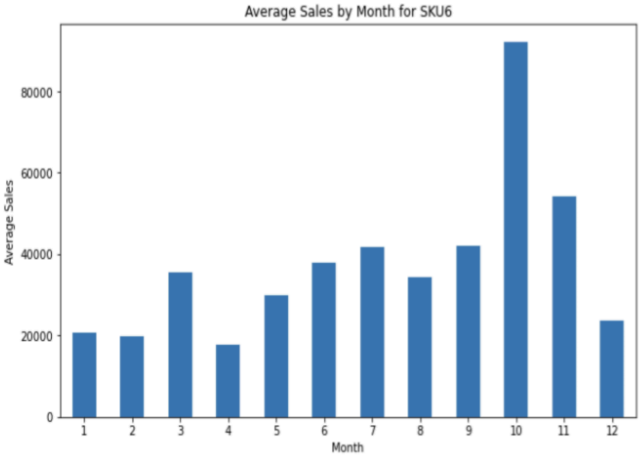
Sku4 sales is high in june august, september and february, and low in december and january. Due to the trend and sales distribution , the products sales are likely affected by seasonality.

Sku5 Monthly Average Sales



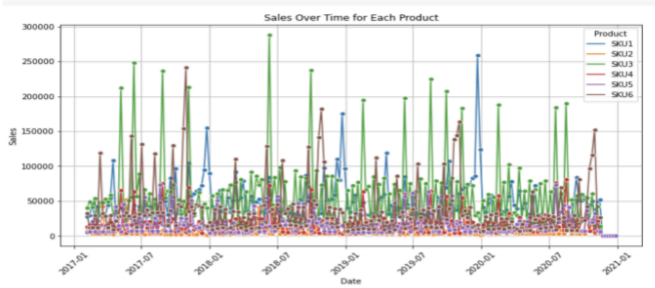
The sales starts going up in may and is at the peak in july and august then it starts dropping in september. This shows a pattern in SKU5 sales and that the product demand is influenced by seasonality.

SKU6 Monthly Average Sales



The average sales of SKU6 is relatively low compared to the other products. In october, the sales spikes but drops in december, january, february and april.This indicates seasonality in product demand, however, marketing strategies should be implemented to improve overall sales.

Sales Overtime and Impact of Price discount on Sales



Sku3, sku1 and Sku4 shows establishment in the market, however sku6, sku5 and sku1 needs more strategies increase sales. Some months, the sales are high and low in some months varying per product this shows seasonality in product demand.

Most of the products sales when the price discount is somewhere between 20% and 50%. sku3 and sku6 had the highest sales when the price discount was around 50%. Sku2 was slightly affected by price discount.

4. **Model Development:**
    - Baseline: Linear Regression.
    - Advanced models: Arima, Random Forest, XGBoost, and LSTM for accurate, explainable forecasts.
        a. **Linear Regression:** In this analysis, we performed a correlation analysis for each SKUs(SKU1, SKU2, SKU3, SKU4, SKU5, SKU6) to evaluate which features are strongly related with the sales of each SKU. Once we identified the key features, we employed a Linear Regression model. For each SKU, we treat the selected features from the correlation matrix as independent variables and the sales as a dependent variable. Model training: we split the data into training and test sets(80%, 20%) respectively, then fit the model on the training data enabling the model to learn the patterns of the data for that particular SKU. Model evaluation: We evaluated each SKU's model using MAE,MSE and R2 score.
        b. **Arima:** In this analysis, we employed the ARIMA (AutoRegressive Integrated Moving Average) model to forecast weekly sales for selected SKUs in the beverage industry. We filtered the sales data for five specific SKUs: SKU1, SKU2, SKU3, SKU5, and SKU6.The sales figures were indexed by date after converting the 'date' column to a datetime format. Any missing sales values were filled with zeros, and rows with zero sales were removed for all SKUs except SKU6 to ensure data integrity.We

conducted the Augmented Dickey-Fuller (ADF) test on the sales data for each SKU to check for stationarity. The results indicated that some time series were non-stationary, prompting the need for differencing. A grid search approach was employed to identify the optimal ARIMA parameters (p, d, q) for each SKU. The data was split into training (before July 2020) and testing (July 2020 onwards) sets.The ARIMA model was fitted to the training data using the best parameters found from the grid search. The model was then used to forecast sales for the test period.

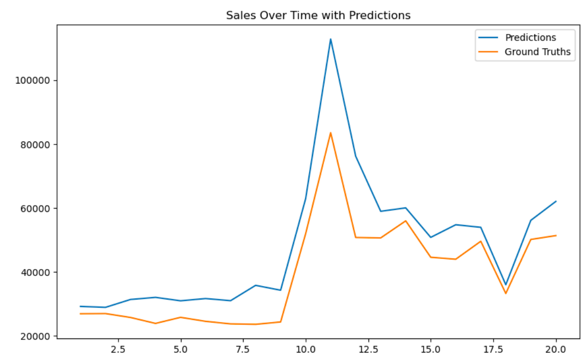
c. **Random Forest:** In this analysis, we employed the Random Forest model to forecast weekly sales for selected SKUs in the beverage industry. We filtered the sales data for six specific SKUs: SKU1 to SKU6. The sales figures were indexed by date after converting the 'date' column to a datetime format. Any missing sales values were filled with zeros, and rows with zero sales were removed for all SKUs except SKU6 to ensure data integrity. Stationarity was tested with Augmented Dickey-Fuller unit root test, and results show that all of them are stationary. We created lag features to capture the temporal dependencies in the sales data. Since we have weekly data, 54 lag variables were generated to account for yearly changes. These features represent the sales values from previous weeks, which help the model learn from past sales trends. For each product, we split the last 40 weeks as testing data. We performed box-cox transformation on the training data, and used the same lambda for the testing data to avoid influence from the testing data. RFE (recursive feature elimination) was used to eliminate the number of features used to 4, Price Discount (%) is always selected, but different lag variables and month were used by different products. 100 estimators were used for the model. After building the model, we fitted it on the testing data, and obtained accuracy by comparing between the predicted and true sales.

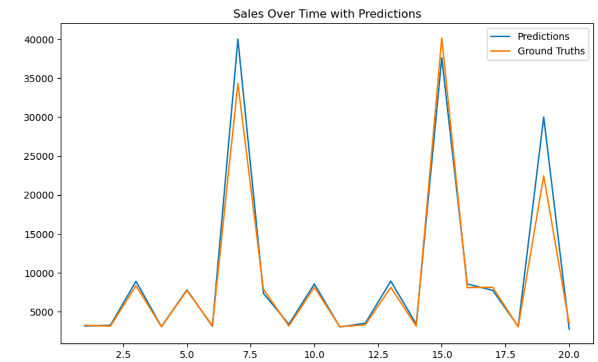
d. **XGBoost:** In this analysis, we employed the XGBoost model to forecast weekly sales for selected SKUs in the beverage industry. We filtered the sales data for five specific SKUs: SKU1, SKU2, SKU3, SKU4,SKU5, and SKU6. The sales figures were indexed by date after converting the 'date' column to a datetime format. Any missing sales values were filled with zeros, and rows with zero sales were removed for all SKUs except SKU6 to ensure data integrity. We created lag features to capture the temporal dependencies in the sales data. Fifty four lagged sales features were generated. These features represent the sales values from previous weeks, which help the model learn from past sales trends. The features selected for the model include: Price Discount, In-Store Promo, Store End Promo, 54 lagged sales. The target variable was the 'Sales' column. The data was

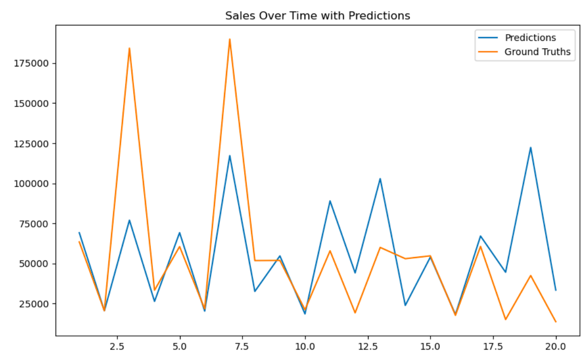
split into training (before July 2020) and testing (July 2020 onwards) sets. The model was trained using the training dataset, fitting the defined features and target variable. The trained XGBoost model was used to make predictions on the test dataset.

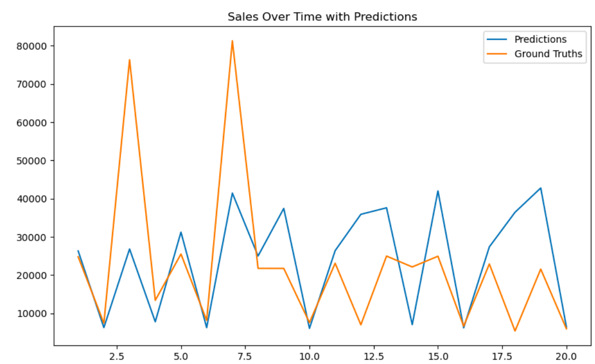
e. **LSTM:** In this analysis, we employed the LSTM model to forecast weekly sales for selected SKUs in the beverage industry. We filtered the sales data for five specific SKUs: SKU1, SKU2, SKU3, SKU4,SKU5, and SKU6. The sales figures were indexed by date after converting the 'date' column to a datetime format. Any missing sales values were filled with zeros, and rows with zero sales were removed for all SKUs except SKU6 to ensure data integrity. LSTMs are sensitive to the scale of input data. Therefore, feature values are normalized using `MinMaxScaler` to ensure they are on a similar scale, typically between 0 and 1. The LSTM model has two layers of LSTM units:The first layer has 100 units, uses ReLU as the activation function, and passes outputs to the next LSTM layer. The second LSTM layer has 50 units, followed by a dropout layer (to reduce overfitting) and batch normalization (to stabilize and speed up training). The model is compiled with the Adam optimizer and uses Mean Squared Error (MSE). The model is trained using early stopping, which helps prevent overfitting by stopping training when the validation loss stops improving. The patience is set to 10 epochs, meaning the model will stop if no improvement is seen for 10 consecutive epochs.A validation split of 20% is used during training to evaluate the model's performance on unseen data. The model outputs a single value, which is the sales prediction.

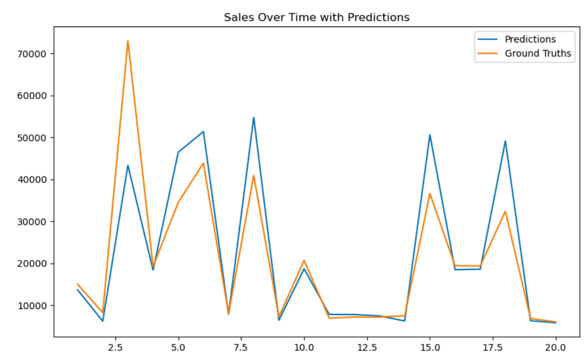
5. **Evaluation & Result:**
    - Used Weighted MAPE & accuracy to measure forecast accuracy.

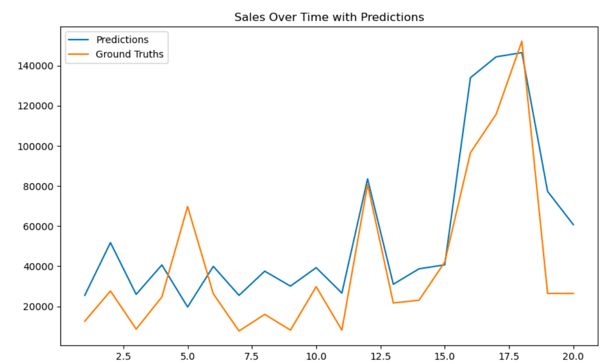












Final Recommendation:

Looking at the performance of the models by checking their MAPE, we decided that Machine Learning models are demonstrating the best performance. Random Forest generally performs better, but for some of the products, XGBoost is better. ARIMA performed well for only SKU 4 compared to other models.