**Sales Forecast Report**

1. **Problem Statement**

In ever-changing competitive market conditions, there is a need to make correct decisions and plans for future events related to business like sales, production, and many more. The effectiveness of a decision taken by business managers is influenced by the accuracy of the models used. Demand is the most important aspect of a business's ability to achieve its objectives. Many decisions in business depend on demand, like production, sales, and staff requirements. Forecasting is necessary for business at both international and domestic levels.

1. **Problem Objective**

Fresh Analytics, a data analytics company, aims to comprehend and predict the demand for various items across restaurants. The primary goal of the project is to determine the sales of items across different restaurants over the years.

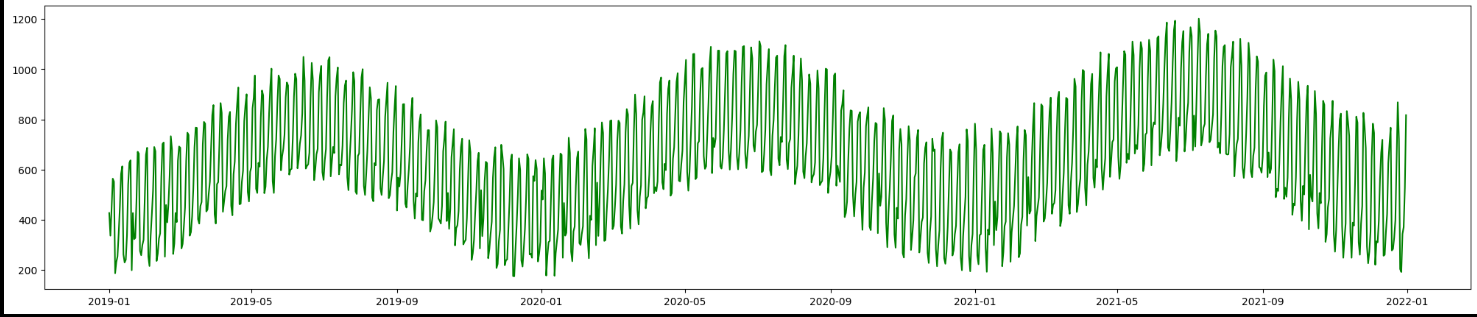
1. **Data Explanation**

The project utilized three datasets: sales, items, and restaurants. The sales dataset contains transaction-level details, including sales volume and store-specific data. The items dataset provides information about the various items, such as item names and categories. The restaurants dataset includes details about restaurant locations and store-specific attributes. These datasets were merged into a single dataset named sales by linking the item\_id from the sales dataset with the id in the items dataset and the store\_id from the sales dataset with the id in the restaurants dataset. The consolidated dataset is ready for analysis, offering a comprehensive view of sales across restaurants.

1. **Exploratory Data Analysis**

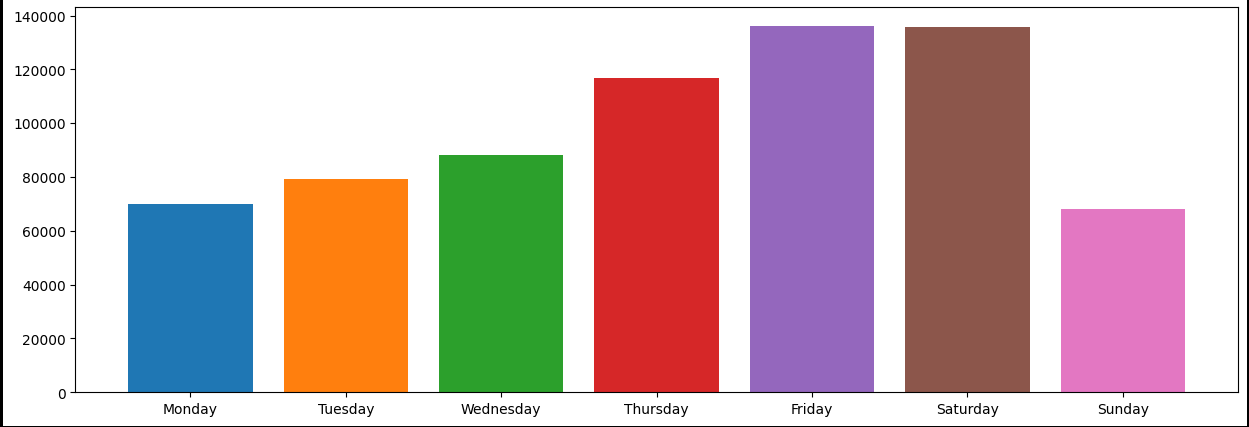
Exploratory Data Analysis was conducted to gain deeper insights into the sales patterns, item popularity, and restaurant performance. This analysis involved examining temporal trends, such as sales variations across days, months, and quarters, as well as exploring item preferences and identifying top-performing restaurants. The findings from this analysis provide a foundation for understanding demand dynamics and support the development of forecasting models.

* 1. **Date-wise sales patterns:**



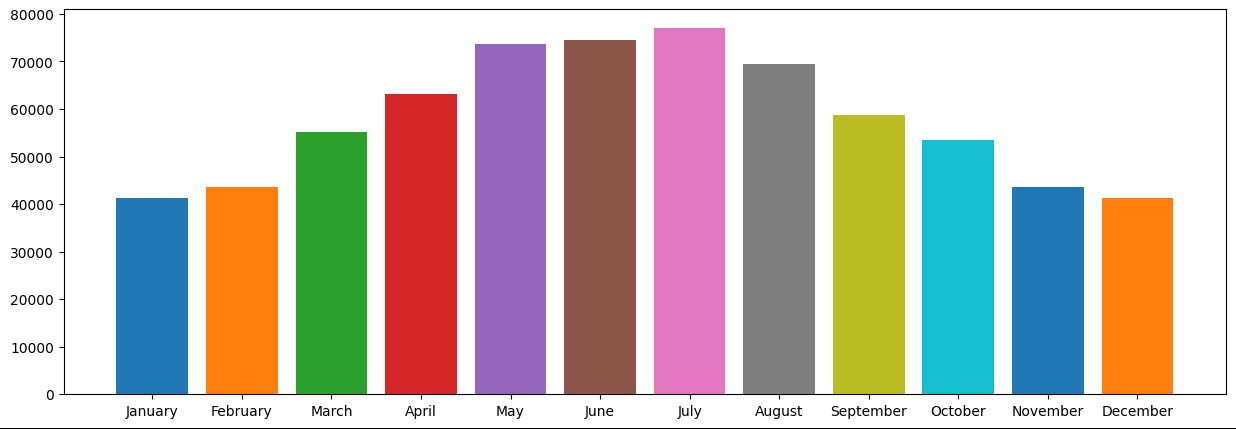
The graph represents the daily sales pattern from 2019 to early 2022, displaying clear seasonality and periodic fluctuations. Sales exhibit consistent peaks and troughs across the years, with higher values around mid-year and lower values toward the year's end. The cyclic nature suggests a strong influence of seasonal trends on the sales.

* 1. **Sales fluctuations across weekdays:**



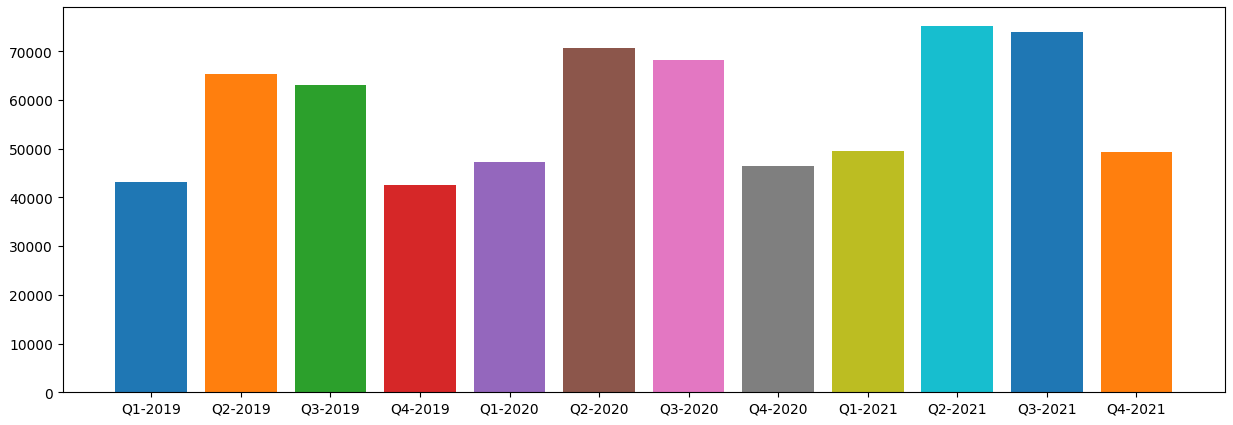
The bar chart illustrates sales fluctuations across the week, with sales increasing steadily from Monday through Thursday. Friday and Saturday show the highest sales, peaking at approximately 140,000 units, indicating strong weekend activity. Conversely, Sunday experiences a sharp drop, suggesting reduced consumer activity on this day. This pattern highlights potential trends for optimizing marketing and inventory strategies.

* 1. **Monthly sales trends:**

****

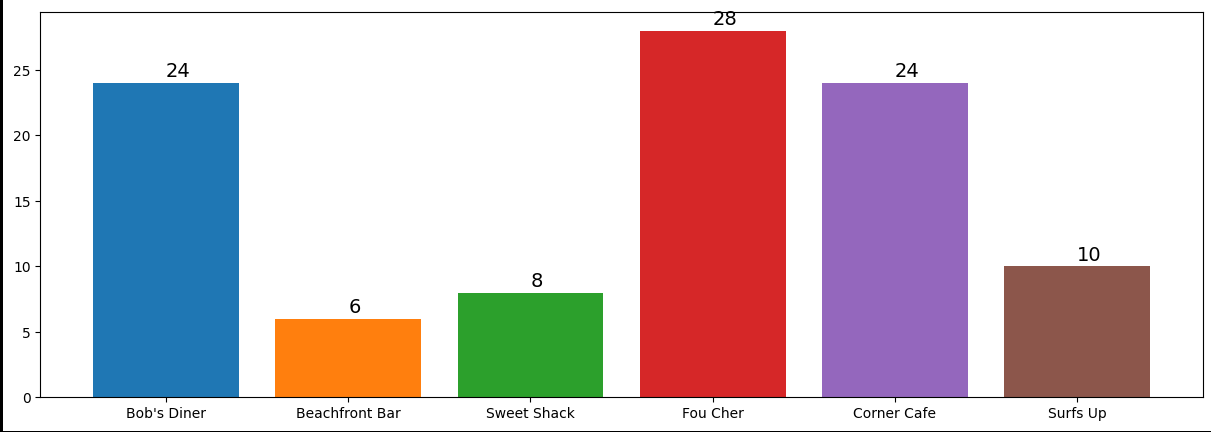
This graph showcases the monthly sales trends, with a gradual increase from January through May, peaking in July at around 80,000 units. Sales slightly decline in August and continue decreasing steadily through the final months of the year, reaching their lowest in November and December. This trend indicates higher consumer activity during the mid-year, possibly due to seasonal factors or holidays, followed by a slowdown toward year-end.

* 1. **Sales distribution across quarters of years:**

****

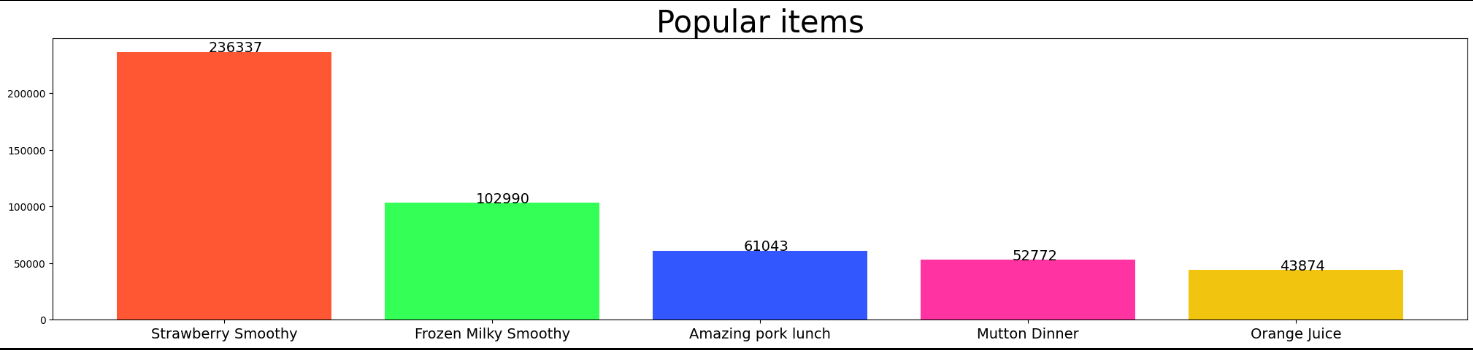
This bar chart represents the sales distribution across quarters from 2019 to 2021. Sales show significant variability, with the second quarters (Q2) of each year consistently achieving higher sales, peaking in Q2-2020 and Q2-2021. In contrast, the fourth quarters (Q4) exhibit relatively lower sales in most years, particularly in Q4-2019 and Q4-2021. This pattern suggests a strong seasonal influence on sales, with mid-year quarters performing better than the year-end quarters.

* 1. **Performance comparison of restaurants:**



The graphical representation compares the performance of five restaurants: Bob's Diner, Beachfront Bar, Sweet Shack, Fou Cher, Corner Cafe, and Surfs Up. Fou Cher leads with the highest performance value of 28, followed by Bob's Diner and Corner Cafe, both at 24. Sweet Shack and Beachfront Bar trail behind with 8 and 6, respectively, while Surfs Up shows moderate performance at 10. The data highlights variations in restaurant performance, suggesting potential areas for deeper investigation into factors influencing these results.

* 1. **Most popular items across the stores:**



This bar graph showcases the popularity of items across stores. "Strawberry Smoothy" is the top-selling item with an impressive count of 236,337, followed by "Frozen Milky Smoothy" at 102,990. "Amazing Pork Lunch" ranks third with 61,043 sales, while "Mutton Dinner" and "Orange Juice" trail with 52,772 and 43,874 sales, respectively. This analysis highlights the dominance of beverages in popularity, particularly smoothies, suggesting a strong consumer preference for refreshing drinks.

* 1. **Forecasting using ML Models**

In this project, multiple machine learning models were implemented to forecast sales based on historical data. The selected models—Linear Regression, Random Forest, and XGBoost, were chosen for their ability to capture linear relationships, complex interactions, and temporal patterns in data.

* + 1. **Linear Regression:**

Linear Regression is a baseline machine learning model used to predict sales by modeling the relationship between input features and the target variable as a linear equation. In this project, the model was trained using historical sales data (X\_train and Y\_train) and applied to predict sales on the test set (X\_test). The predictions (lr\_pred) serve as a benchmark for evaluating the performance of more complex models.

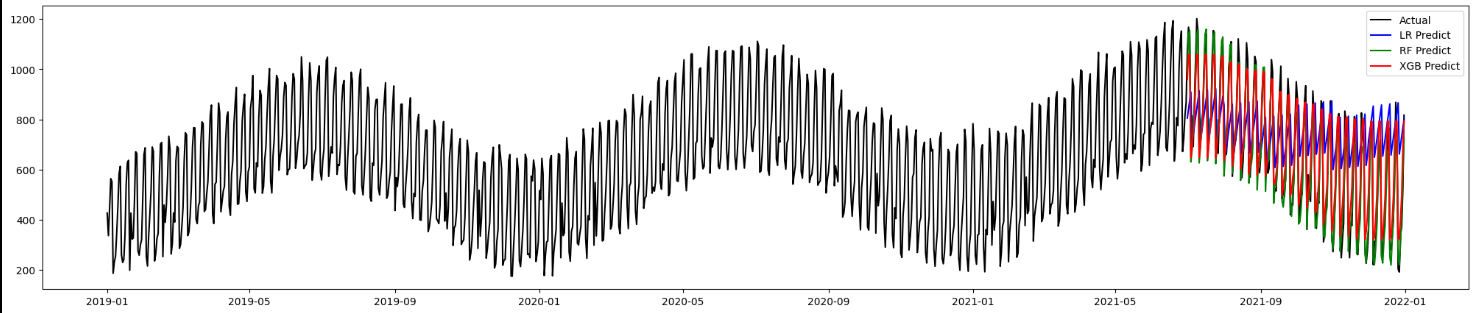
* + 1. **Random Forest:**

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. For this project, a Random Forest Regressor with 100 estimators and a maximum depth of 21 was trained on the sales data (X\_train and Y\_train). The model captured complex, non-linear relationships in the data and generated predictions (rf\_pred) for the test set (X\_test). Its robustness makes it particularly effective for forecasting tasks.

* + 1. **XGBoost:**

XGBoost is a powerful gradient boosting algorithm designed to optimize performance and efficiency. In this project, the XGBoost Regressor was configured with hyperparameters such as 1000 estimators, a maximum depth of 3, and a learning rate of 0.01. The model was trained on X\_train and Y\_train while using early stopping rounds to prevent overfitting. XGBoost's ability to handle non-linear patterns and feature interactions makes it highly effective for sales forecasting, providing precise predictions on the test set (X\_test).

* 1. **Results Comparison**



Random Forest emerges as the most accurate model with the lowest RMSE (60.09) and MAE (47.77), indicating minimal errors in its predictions. It also has the highest R2R^2R2 score (94.41), demonstrating that it captures the variability in the sales data exceptionally well. XGBoost, while slightly less accurate, performs robustly with an RMSE of 71.18 and MAE of 58.45, and a high R2R^2R2 score of 92.15. These results show that XGBoost effectively models the data but is slightly outperformed by Random Forest.

Linear Regression, however, falls significantly behind with a much higher RMSE (244.21) and MAE (196.04), as well as a very low R2R^2R2 score (7.65), indicating it fails to explain the variability in the data effectively. This comparison highlights that Random Forest is the most suitable model for forecasting sales in this scenario, followed closely by XGBoost, with Linear Regression being inadequate for this task.

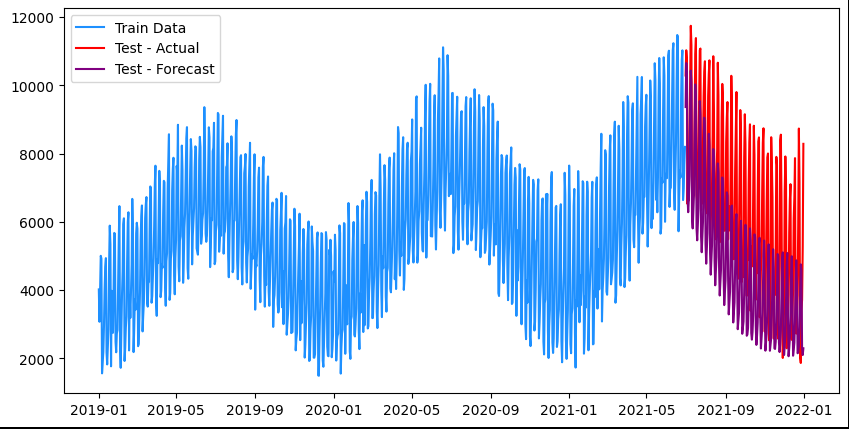
1. **Forecasting with DL Model**

Deep learning models are well-suited for capturing intricate patterns and dependencies in sequential data. In this project, an LSTM (Long Short-Term Memory) network was employed to leverage its ability to analyze temporal trends and predict future sales with high accuracy.

* 1. **LSTM**

The LSTM (Long Short-Term Memory) model is designed for time series forecasting in this project. It consists of an LSTM layer with 100 units and ReLU activation, which processes sequential data of specified input shape (length and number of features). A Dense layer is added as the output layer to predict a single continuous value. The model is compiled using the Adam optimizer and Mean Squared Error (MSE) as the loss function. Early stopping is employed to halt training if no improvement is observed in validation loss over consecutive epochs, preventing overfitting. The model is trained using a TimeseriesGenerator for the test data.

* 1. **Performance**



The plot highlights the strong performance of the LSTM model in forecasting future sales. The training data (blue line) shows clear seasonal patterns, which the model successfully captures. During the test period, the predicted sales (purple line) align closely with the actual sales (red line), demonstrating the model's ability to effectively learn and replicate both the trends and seasonality in the data. While minor deviations exist, these are typical in real-world forecasting scenarios and do not significantly impact the overall accuracy. The Mean Absolute Percentage Error (MAPE) of 37.86% indicates the model provides reliable predictions and performs well for a dataset with such inherent complexity. This result highlights the LSTM model's potential as a robust tool for time series forecasting.

1. **Conclusion**

In this analysis, various machine learning and deep learning models were used to forecast sales, providing insights into demand patterns. Among the models, Random Forest delivered the most accurate results, with the lowest RMSE and MAE and the highest R² score, effectively capturing the data's complexity. XGBoost also performed well, while Linear Regression fell short due to its inability to model non-linear patterns.

The LSTM model excelled in capturing temporal trends and seasonality, aligning closely with actual sales during the test period. With a MAPE of 37.86%, it demonstrated reliable forecasting performance for complex datasets. Overall, the study highlights the strength of advanced models like Random Forest and LSTM in supporting accurate, data-driven decision-making for optimizing sales and resource planning in dynamic markets.