RETAIL INVENTORY ANALYSIS AND DEMAND FORECASTING

PDS PROJECT TEAM 8

Group members:

YV SIVA SURYA	S20200010245
SURAPUCHETTY ABHILASH	S20200010216
P SATYA SAI RAMESH	S20200010162
N NAGANNA DORA SASTRY	S20200020298

Retail Inventory Analysis and Demand Forecasting

YV SIVA SURYA
Department of CSE, IIIT SRICITY
Email: venkatasivasurya.y22@iiits.in
Roll No: S20220010245

SURAPUCHETTY ABHILASH
Department of CSE, IIIT SRICITY
Email: abhilash.s22@iiits.in
Roll No: S2022002010216

P SATYA SAI RAMESH
Department of CSE, IIIT SRICITY
Email: satyasairamesh.p22@iiits.in
Roll No: S20220010162

N NAGANNA DORA SASTRY
Department of ECE, IIIT SRICITY
Email: nagannadorasastry.n22@iiits.in
Roll No: S20220020298

Abstract

Retail businesses rely on efficient inventory management to balance supply and demand, reduce losses, and optimize sales.

This study examines a retail store inventory dataset containing sales data, inventory levels, demand forecasts, external factors such as weather and seasonal trends. Through comprehensive data preprocessing, we address missing values, remove outliers using the IQR method, and analyze demand distribution. Exploratory data analysis (EDA) reveals key insights into sales trends, product demand fluctuations, and pricing effects.

Additionally, machine learning models, are employed for demand forecasting, aiming to enhance inventory decision-making. The findings from this study provide actionable insights for businesses to improve stock management, minimize wastage, and maximize revenue.

1. Introduction

Effective inventory management is a cornerstone of successful retail operations. Businesses must balance supply and demand to prevent stockouts, overstocking, and financial losses. Traditional inventory management strategies often rely on static rules or past sales trends, which may not effectively capture dynamic market conditions.

This study focuses on analyzing a retail inventory dataset that includes historical sales, stock levels, pricing, discounts, competitor pricing, and external influences such as weather conditions and seasonality. The objective is to identify patterns that influence demand and develop predictive models to optimize inventory management.

To achieve this, we conduct thorough data preprocessing, including handling missing values and outliers, followed by exploratory data analysis (EDA) to extract meaningful insights. Finally, we implement machine learning techniques, to forecast demand and enhance stock replenishment strategies. By leveraging data-driven decision-making, retailers can improve efficiency, reduce losses, and better respond to consumer demand.

2. Problem Statement

Retailers often face challenges in maintaining the right stock levels due to unpredictable demand fluctuations. Overstocking leads to increased holding costs and potential product wastage, while stockouts result in lost sales and dissatisfied customers. Traditional inventory management approaches may not adequately capture the complex relationships between sales, pricing, external factors, and seasonality.

This study aims to bridge this gap by analyzing a retail store inventory dataset and applying predictive modeling to improve demand forecasting. By integrating machine learning techniques we seek to develop a robust demand prediction model. The goal is to provide businesses with actionable insights to optimize stock levels, minimize losses, and enhance overall retail efficiency.

3. Methodology

The analysis began with the installation of necessary libraries, including missingno for visualizing missing data and lightgbm for potential machine learning tasks. The dataset was then loaded using pandas, and

its structure was explored to gain an initial understanding.

For preprocessing, essential libraries such as numpy and pandas were imported for data manipulation, while seaborn and matplotlib were used for visualization. Additionally, sklearn and boruta were incorporated to facilitate machine learning tasks and feature selection.

The exploration phase involved generating basic information and summary statistics to assess the dataset's characteristics, enabling a comprehensive understanding of its composition and potential preprocessing needs.

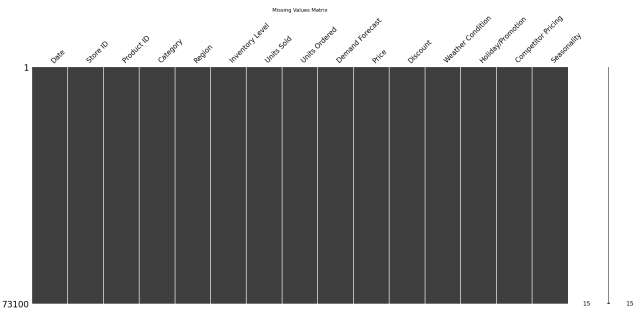
3a. Data Initial Preprocessing

3a.i Initial features and data types :

Feature	Feature
Date (object)	Units Sold (int64)
Store ID (object)	Units Ordered (int64)
Product ID (object)	Demand Forecast (float64)
Category (object)	Price (float64)
Region (object)	Discount (int64)
Inventory Level (int64)	Seasonality (object)
Competitor Pricing (float64)	Weather Condition (object)
	Holiday/Promotion (int64)

Statistic	Inventory Level	Units Sold	Units Ordered	Demand Forecast	Price	Discount	Holiday/Promotion	Competit
Count	73100	73100	73100	73100	73100	73100	73100	73100
Mean	274.47	136.46	110.00	141.49	55.14	10.01	0.50	55.15
Std	129.95	108.92	52.28	109.25	26.02	7.08	0.50	26.19
Min	50.00	0.00	20.00	-9.99	10.00	0.00	0.00	5.03
25%	162.00	49.00	65.00	53.67	32.65	5.00	0.00	32.68
50%	273.00	107.00	110.00	113.02	55.05	10.00	0.00	55.01
75%	387.00	203.00	155.00	208.05	77.86	15.00	1.00	77.82
Max	500.00	499.00	200.00	518.55	100.00	20.00	1.00	104.94

Statistics of our data



Heat map of null values

Observations:

- As we see there are no null values in the data
- These are numerical features
  - Inventory Level, Units Sold, Units Ordered, Price, Discount, Holiday/Promotion, Competitor Pricing
- These are categorical features
  - Store ID(5 values), Product ID(20 values), Category(5 values),Region(4 values), Weather Condition(4 values), Seasonality(4 values)

3b. Exploratory Data Analysis

3b.i Understanding the data distribution:

Visualisation of data distribution

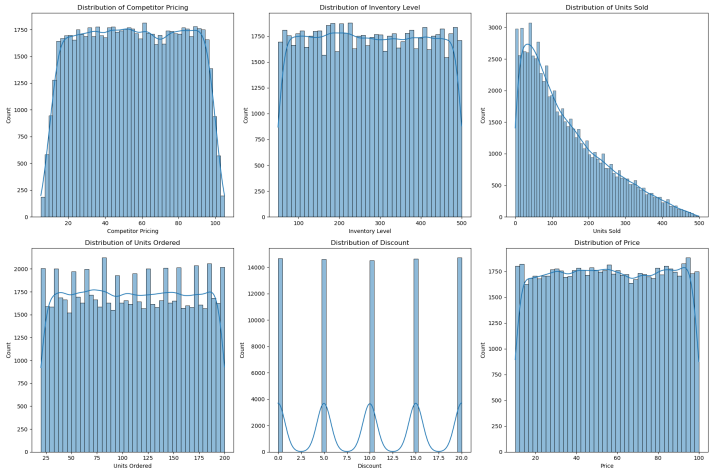


Fig 1.1

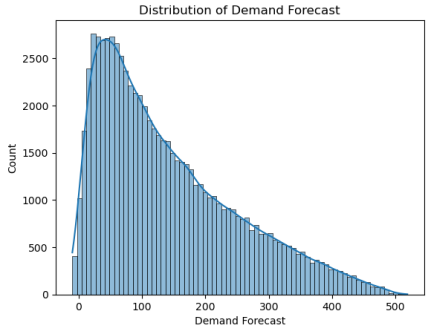


Fig 1.2

Observations:

- Based on the data distributions we see units sold and demand forecast is right skewed in case if there are outliers we should use IQR for filtering them
- And also they had a similar distribution there is very good chance of them being very related

3b.ii Finding outliers in the data:

Visualisation :

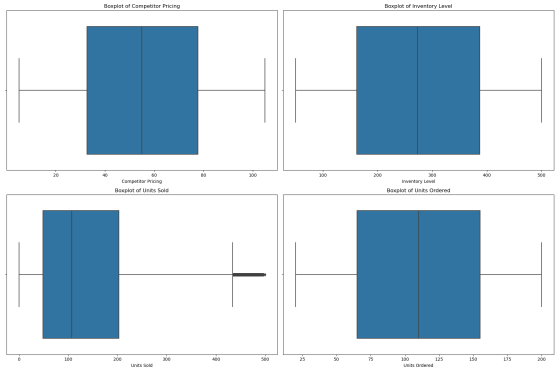


Fig 2.1

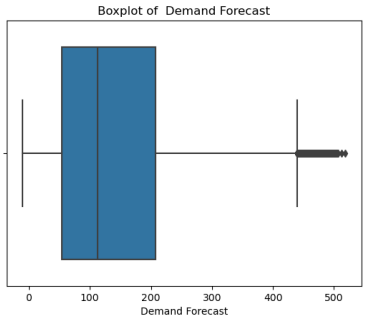


Fig 2.2

Observations:

- There are outliers in the units sold and forecast demand

3b.iii Pattern analysis:

Correlation Analysis:

- Identified the patterns and trends in the dataset, such as correlations between numerical features like Inventory Level, Units Sold, Units Ordered, Price, Discount, Holiday/Promotion, Competitor Pricing

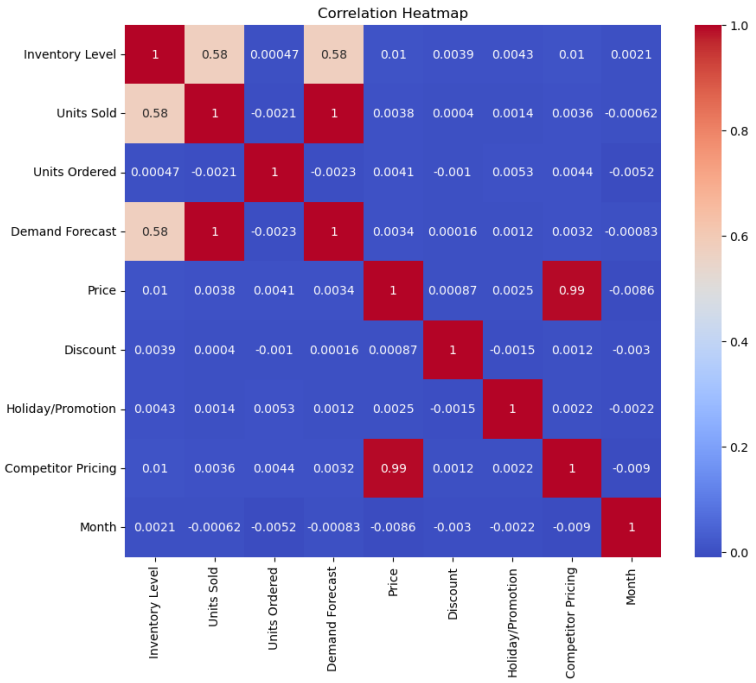


Fig 3.1

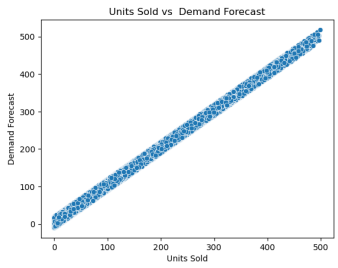


Fig 3.2

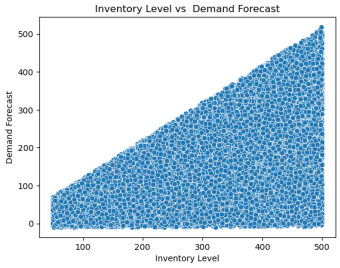


Fig 3.3

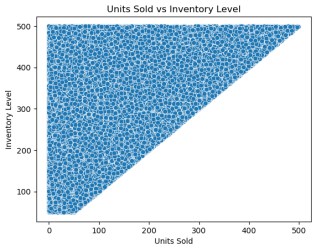


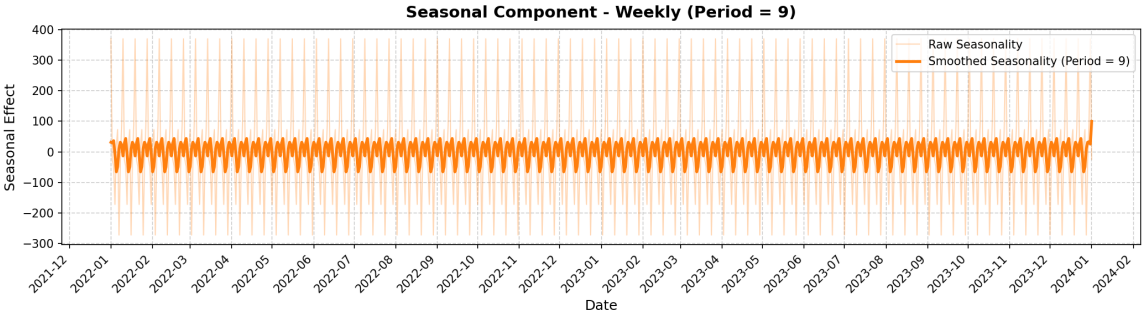
Fig 3.4

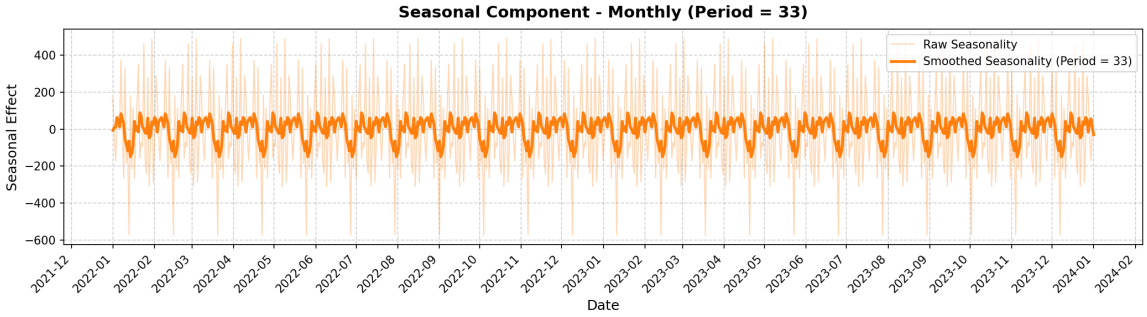
Observations:

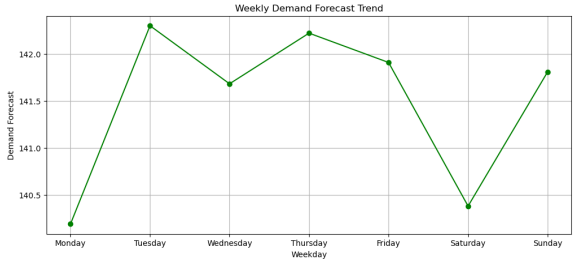
- Based on these correlation matrix the most corelated to data are inventory level and units sold
- The price and competitor price are more correlated

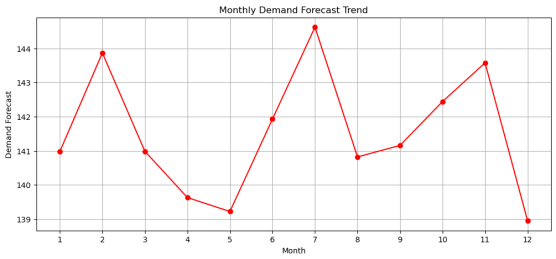
3b.iv Seasonal Trends:

- Our seasonality analysis examined the **weekly and monthly demand forecast**, identifying **seasonal components, trends, and residuals** to understand fluctuations over time. By visualizing the **average demand forecast** at both levels, we detected recurring patterns that influence demand variations.









Observations:

- The demand forecast has a hidden weekly and monthly component
- At Monday and Saturday the demand is low and all the other days its peak in a week
- In all the months demand forecast reaches peek at February, July, November on average

3b.v Categorial trends:

Visualisation :

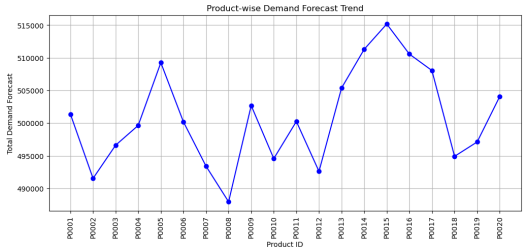


Fig 4.1

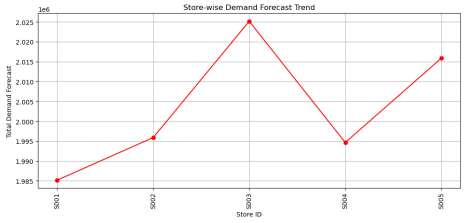


Fig 4.2

4. Data processing based on EDA

Removing outliers:

From the boxplots of numerical features we had found out both demand forecast and units sold has considerable outliers

As they are right skewed we used IQR to remove the outliers

5. Feature engineering and further processing:

5a. Features added based on EDA covariance analysis:

Based on the correlated features we came up with some more new features as defined below:

$$\text{Sales Velocity} = \frac{\text{Units Sold}}{\text{Inventory Level} + 1} \qquad \text{Effective Price} = \text{Price} \times \left(1 - \frac{\text{Discount}}{100}\right) \qquad \text{Competitive Price Gap} = \text{Competitor Pricing} - \text{Effective Price}$$

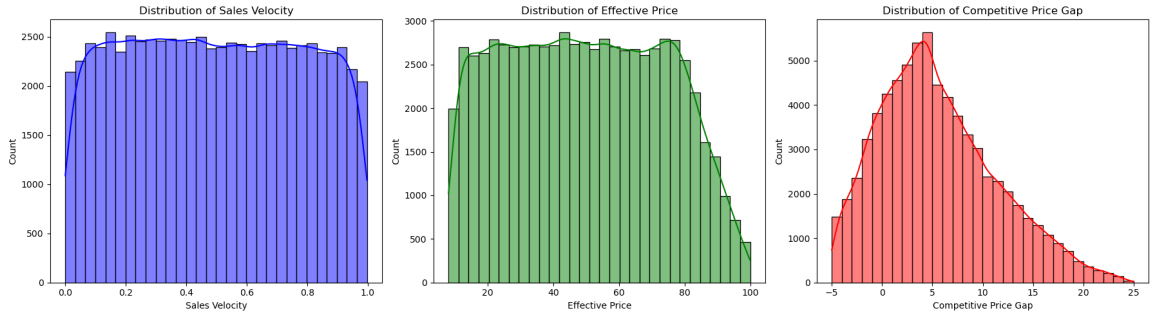
Further analysis on these features:

After removing null values we had visualised and crosschecked data distribution how much these new features are contributing

Statistics

Feature Name	Count	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
Sales Velocity	71,716	0.495	0.284	0.000	0.250	0.493	0.740	0.998
Effective Price	71,716	49.63	23.81	8.008	29.214	49.212	69.64	99.98
Competitive Price Gap	71,716	5.532	5.833	-5.00	1.200	4.708	9.166	24.972

Distribution:



Scatter plots to identify how are they related:

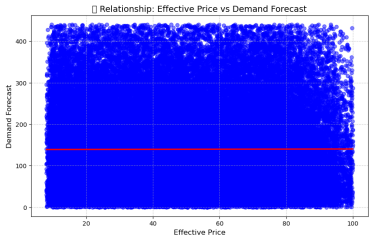


Fig 5.1

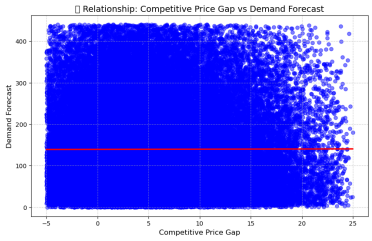


Fig 5.2

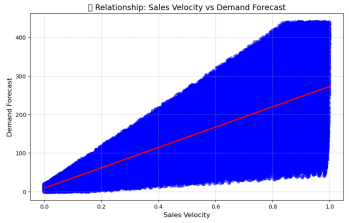


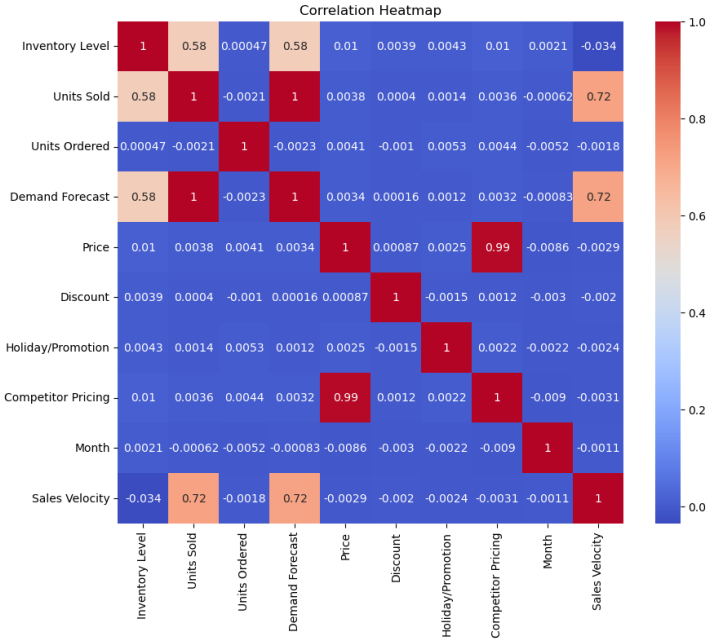
Fig 5.3

Correlation analysis:

Feature	Sales Velocity	Competitive Price Gap	Effective Price
Demand Forecast	0.7157	0.0025	0.0029

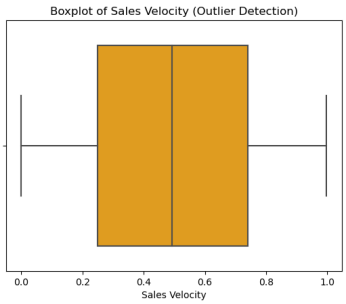
Observations:

- Based on the scatter plots we find only the Sales Velocity is the useful feature



Processing Sales velocity feature:

Finding outliers: Visualising Sales velocity:



There are no outliers in sales velocity

5b. Features added based on seasonal trend:

Based on the seasonality trends like the weekly trend we came up with some more new features as defined below:

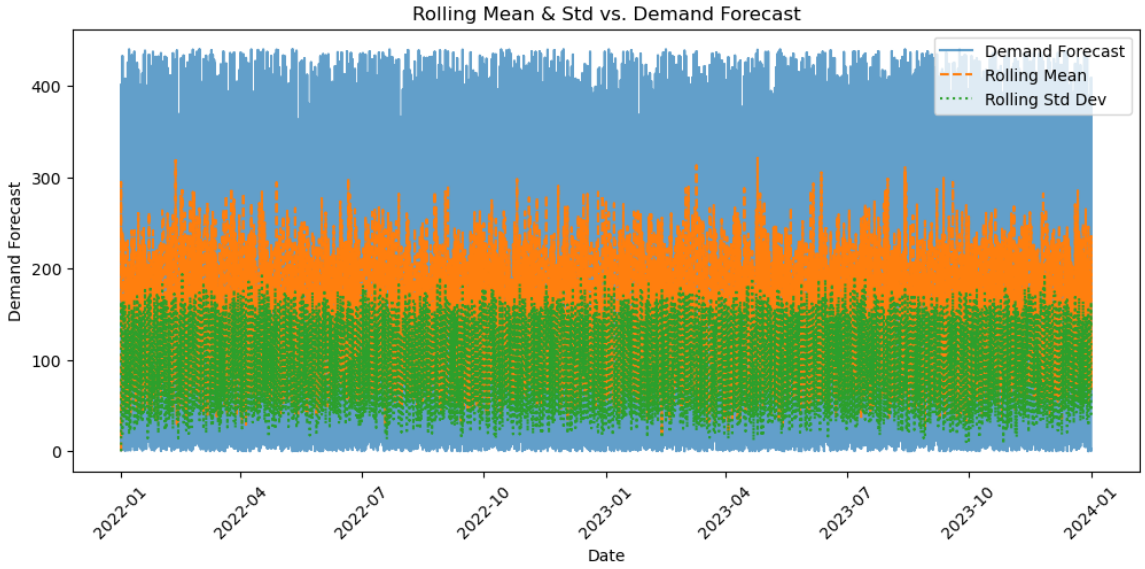
$$\text{Rolling\_Std\_7}_t = \sqrt{\frac{1}{7}\sum_{i=t-6}^{t}(\text{Demand Forecast}_i - \text{Rolling\_Mean\_7}_t)^2}$$

$$\text{Rolling\_Mean\_7}_t = \frac{1}{7}\sum_{i=t-6}^{t}\text{Demand Forecast}_i$$

Further analysis on these features:

After removing null values we had visualised and crosschecked data distribution how much these new features are contributing

Distribution:



Scatter plots to identify how are they related:

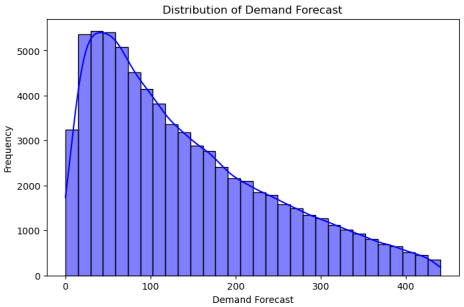


Fig 5.1

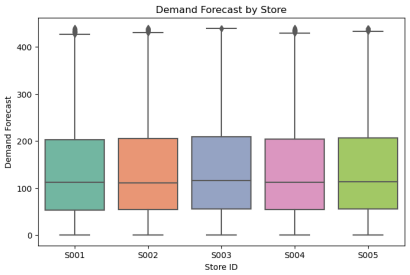
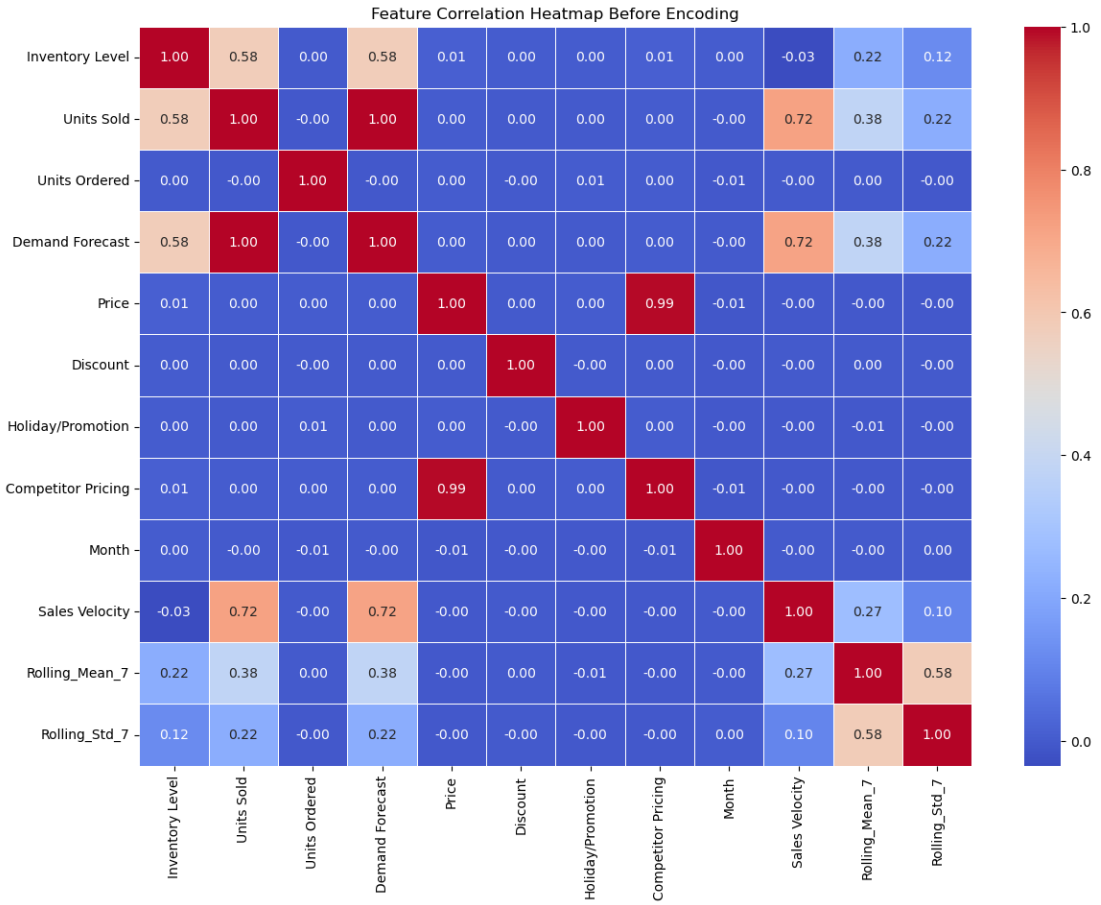


Fig 5.2

6. Final analysis and preprocessing with all existing and new features included

6a. Final Complete Correlation analysis:



6b. Variance analysis:

Feature	Original Variance	MinMax Variance	Z-score Variance
Inventory Level	16,592.79	0.0819	1.0
Units Sold	10,907.94	0.0543	1.0
Units Ordered	2,733.20	0.0844	1.0
Price	677.49	0.0836	1.0
Discount	50.19	0.1255	1.0
Holiday/Promotion	0.25	0.25	1.0
Competitor Pricing	686.44	0.0688	1.0
Month	11.92	0.0985	1.0
Sales Velocity	0.08	0.0810	1.0
Rolling_Mean_7	1,553.78	0.0151	1.0
Rolling_Std_7	905.52	0.0241	1.0

6c. Encoding of categorical variables:

The categorical variables are converted to boolean using LabelEncoder from sklerarn

Here's a summary of how the categorical variables were encoded and their transformations:

Column Name	Encoding Method
Store ID	One-Hot Encoding
Category	One-Hot Encoding
Region	One-Hot Encoding
Weather Condition	One-Hot Encoding
Seasonality	One-Hot Encoding
Weekday	One-Hot Encoding
Product ID	Label Encoding

6d. Scaling data:

Based on the variance of numerical features we used MinMax scaling so it r educes variance significantly, ensuring all features are within a [0,1] range

7a. Ml Modelling:

7a.i Linear Regression Model:

Due to the more correlated variables in all features like Units Sold, Inventory level, Sales Velocity .Hence, we employed a multi linear regressor model to predict demand forecast

7a.ii Random Forest Regressor Model:

Alongside more correlated variables to compensate moderately correlated variables which may not essentially linear like Rolling mean ,Rolling Std and also to identify categorial relationships(encoded) .Hence, we employed Random Forest Regressor model to predict demand forecast

7a.iii Bayesian Regressor Model:

Bayesian Regressor Model helps to find Seasonality like weekly yearly season patterns and categorial relationships with combining more correlated variables in all features like Units Sold, Inventory level, Sales Velocity

7a.iv GradientBoost on Decision Tree Regressessor Model:

There are correlated features and moderately correlated variables equally and has seasonal patterns, and category trends.A GradientBoost can handle these non linear relations repeating patterns .Hence, we employed GradientBoost on Decision Tree Regressessor Model to predict demand forecast.

7b. ML Results:

Model	MAE	MSE	R^2 Score
Random Forest	7.64739	79.92322	0.99268
Linear Regression	18.65922	573.83418	0.94746
Gradient Boosting	7.82499	84.57915	0.99226
Bayesian Regression	7.46253	74.37914	0.99319

Based on the all the model results best fit is Bayesian Regressor Model here is the QQplot and scatter plot of the predictions of the bayesian regressor wrt ground truth

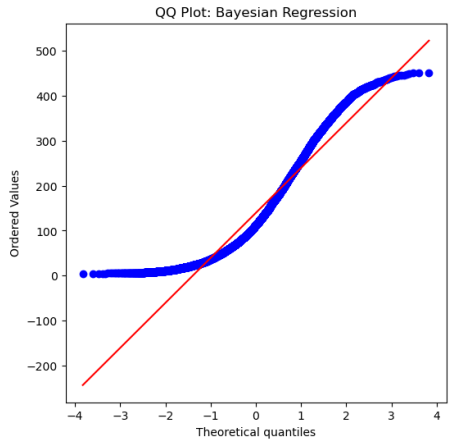


Fig 6.1

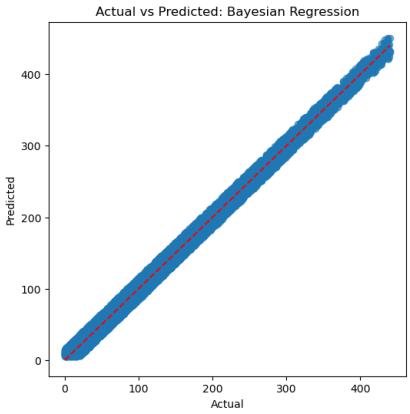


Fig 6.2

8. Conclusion

In this project, we successfully applied demand forecasting techniques and inventory analysis to optimize retail operations. By implementing machine learning models and inventory management strategies, we were able to predict demand with reasonable accuracy and ensure optimal stock levels, reducing both overstocking and stockouts. The integration of a user-friendly dashboard further enhanced the decision-making process, enabling better inventory control.

The work carried out in this project has significant potential for real-world applications in retail environments, where accurate demand forecasting can drastically improve operational efficiency. However, there is always room for improvement, particularly in the areas of model accuracy and adaptability to dynamic market conditions.

For future work, we plan to explore more advanced forecasting techniques, incorporate real-time data, and refine the dashboard for even better user interaction. Additionally, we aim to test our model in various retail sectors to validate its scalability and effectiveness.

9. Team Contribution

i.Naganna : Preproceesed data by removing null values, duplicates, and outliers using the IQR method. Analyzed data distributions and trends, focusing on average values for categorical features. Built a Random Forest regressor model to predict demand forecast.

ii.Ramesh: Correlation analysis of the inital numerical fetaures ,through Sesonal decomposition found the weeky and monthly repeating trends and residual fluctuations and detected month wise outliers .Built a Linear Regression Model

iii.Siva Surya :To improve demand forecasting and sales predictions, we added features like Sales Velocity , Effective Price and Competitor Price Gap to better capture price relationships. Since the data was left-skewed, we used Variance analysis to ensure these features were relevant. To balance correlations and seasonal trends, we chose a Gradient Boosting Decision Tree Regressor for its ability to handle complex patterns.

iv.Abhilash: Added new features like Rolling Mean,Rolling Std Came up With MinMax Scaling based on the variance analysis. added a closing correlation analysis with all features together. Built a Bayesian

Regressor Model to predict demand forecast.

Our team collaborated effectively by holding regular meetings to discuss progress, challenges, and improvements. Each of us given our share of efforts in project—from data collection to implementation—were completed successfully. Each of us came up with our own proposals in EDA , Feature Engineering,ML models based an mutual analysis and feedback. We maintained continuous communication and shared feedback to refine our work and achieve the best results.