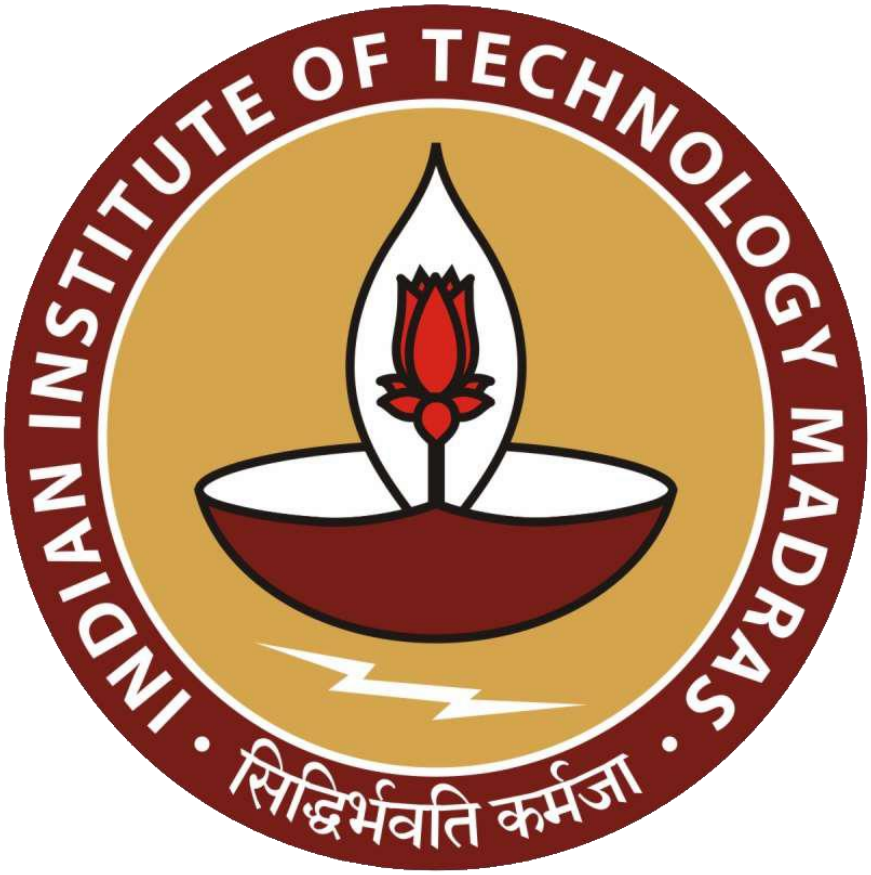
Optimising Operations: Strategies to Overcome Insufficient Workforce, Superfluous Inventory and Inadequate Clientele of a Wholesale Company.

Final Report for BDM Capstone Project

Submitted By:

Ikshwaku Tiwari

21f1003999



IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

Executive Summary

Nalanda Dresses, established in 1970 by Mr. Subodh D. Shah, is a wholesaling company that manufactures and distributes only girl's apparel. The primary problems faced by the organisation included overflowing inventory due to unplanned production, acquiring new skilled labour and retaining the current employees, and no local presence in the market.

The data collected consisted of the sales figures, provided by the company itself, which is the basis for the analysis conducted in this report and the qualitative data by interacting with the owners. The sales data consisted of Sale Figures, Items Sold, Date and the Locations of the customer companies.

Using all the available data Machine Learning models, specifically Linear Regression and XG-Boost were created that were able to analyse the existing features and create a sales prediction model with around 97 per cent accuracy. The performance metrics used were regression metrics like Mean Square Error and R2 score, and the plots created highlighted the model efficiency and the important features that contributed towards model building.

The other part of the analysis included geographical visualisations that showed what parts of India the company cater to and how much hypothetical cost is incurred while travelling to these distant parts. The results highlighted the customer companies belonging to the southern part of India as 90 per cent of the buyers.

Final recommendations include the way inventory levels can be optimised based on the Machine Learning model, and creating marketing campaigns to induce a local presence.

Detailed Explanation of Analysis Methods:

Part 2.1: Demand Forecasting Based on Sales Prediction for Inventory Optimisation:

The data provided by the business organisation included the following features/columns:

Date	Company Name	City/Town	Sales	Items

As mentioned in the previous reports, the company has no viable records or patterns for inventory management. It relies on the traditional approach of 'producing when demanded'. Due to this approach, there are often cases of excess inventory, which leads to an increase in labour and storage costs.

To tackle the problem of inventory pile-up, demand forecasting using a sales prediction approach is chosen. The objective of the analysis is to predict future sales, considering sales as the target variable. Since the analysis involves a machine-learning approach, Python libraries like Pandas, NumPy, Seaborn, Matplotlib and Scikit-Learn are used.

Steps involved in the end-to-end ML approach:

1. Data Preprocessing and Exploratory Data Analysis:
    - Using the Pandas library in Python, the overall structure, unique values, missing data, and value counts of the important features are explored.
    - The Sales feature had a categorical data type due to the '₹' symbol. This is converted to a float datatype for analysis.
    - Descriptive statistics numerical and visual, like a Box-Plot for Sales Distribution, a Histogram for Items Sold, and a Bar-Chart for cumulative transactions for each city/town.
    - The date column, initially a categorical datatype, is converted to a DateTime datatype, so that features like day, month, and year are extracted, which contribute to the final model building.
    - Using these features and the Items feature, new interactive features are created like Items_dayofweek and Items_month.
    - As the sales figures were considerably larger compared to the rest of the data as well as slightly right-skewed, a log transformation is applied to sales.
    - With the target variable separated from the rest of the dataset, a train-test split is applied to the dataset with 80 per cent of training data and 20 per cent of testing data.
    - Lastly, all the numerical features are scaled using the Standard Scaler, an API that standardizes features by removing the mean and scaling to unit variance, using the formula $z=(X-\mu)/\sigma$, where μ is the mean and σ is the standard deviation, and the categorical features are encoded using One-Hot Encoder API, that transforms categorical variables into a binary matrix, where each category is represented by a unique binary column.

2. Model Building for Sales Prediction:
    - Predicting sales using existing sales comes under supervised machine learning, specifically regression, as we are predicting a continuous variable.
    - A dictionary with the most efficient regression model APIs in the Scikit-Learn library is created.
    - The training data set is fitted upon each model, and assessed using the MSE (Mean Squared Error) and the R2 score of each model.
    - The best 3 models based on the performance metrics are chosen for further analysis.
    - Checking for 'overfitting' for the best model, as the training and testing data scores might be different which will hamper further analysis.

3. Model Evaluation:
    - Hyperparameter Tuning is performed for the Linear Regression and XG-Boost Models as they give the best performance.
    - Two approaches have been taken to evaluate the performance of the selected models. The first involves not capping the outliers for Linear Regression and the second approach involves capping the outliers with Winsorization.
    - Plots like residual plots and feature importance plots have been created to compare and visualize the model performance.
    - Final Plots include the Actual vs Predicted sales figures plot, which helps predict future sales of the company.

Mathematical Explanation of the ML Models Used:

1. Linear Regression:
   Linear regression is the simplest of all regression models. It comes under supervised learning as the model uses available predictions to learn and then predicts a continuous variable based on the predictions. The objective function of the Linear Regression is as follows:

   $$\log(1+\text{Sales}) = \beta 0+\beta 1X1+\beta 2X2+\cdots+\beta pXp+\epsilon, \text{ (After applying log)}$$

   And the loss function (MSE) is:

   $$min_{\beta}\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2$$

2. XG Boost:
   XG Boost is an ensemble model (created by constructing multiple decision trees together) that uses the gradient boosting algorithm. XG Boost is widely used in regression and classification tasks and also comes under supervised learning. XG Boost is much more complex and robust compared to Linear Regression, as it consists of multiple parameters that take into account factors like non-linearity, features on different scales, multiple parameters to work with etc. The objective function of XG Boost is as follows:

   $$\text{Objective} = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

   Where,
   $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda|w|^2$ is the regularization parameter.

Part 2.2: Workforce Cost Analysis:

The qualitative data that was collected by interacting with the owners consisted of the labour data, i.e. the total number of employees and their job distribution. Also, the kinds of dress material used to manufacture dresses and their selling price range.

The analysis conducted using this data had the following objective: To estimate the workforce cost per fabric.

After combining the two datasets the mid-price of the dress materials was calculated:

$$\text{Mid Price} = \frac{\text{Min Price} + \text{Max Price}}{2}$$

Then the number of employees allocated per fabric was calculated:

$$\text{Employees per Fabric} = \frac{\text{No of Employees}}{\text{Total Number of Fabrics}}$$

And finally, the estimated cost is calculated:

$$\text{Estimated Cost} = \text{Employees Allocated} \times \text{Mid Price}$$

This analysis is accompanied by a bar chart that helps visualise the cost-heavy fabrics.

Part 2.3: Analysing the Geographical Locations and the cost incurred for transport.

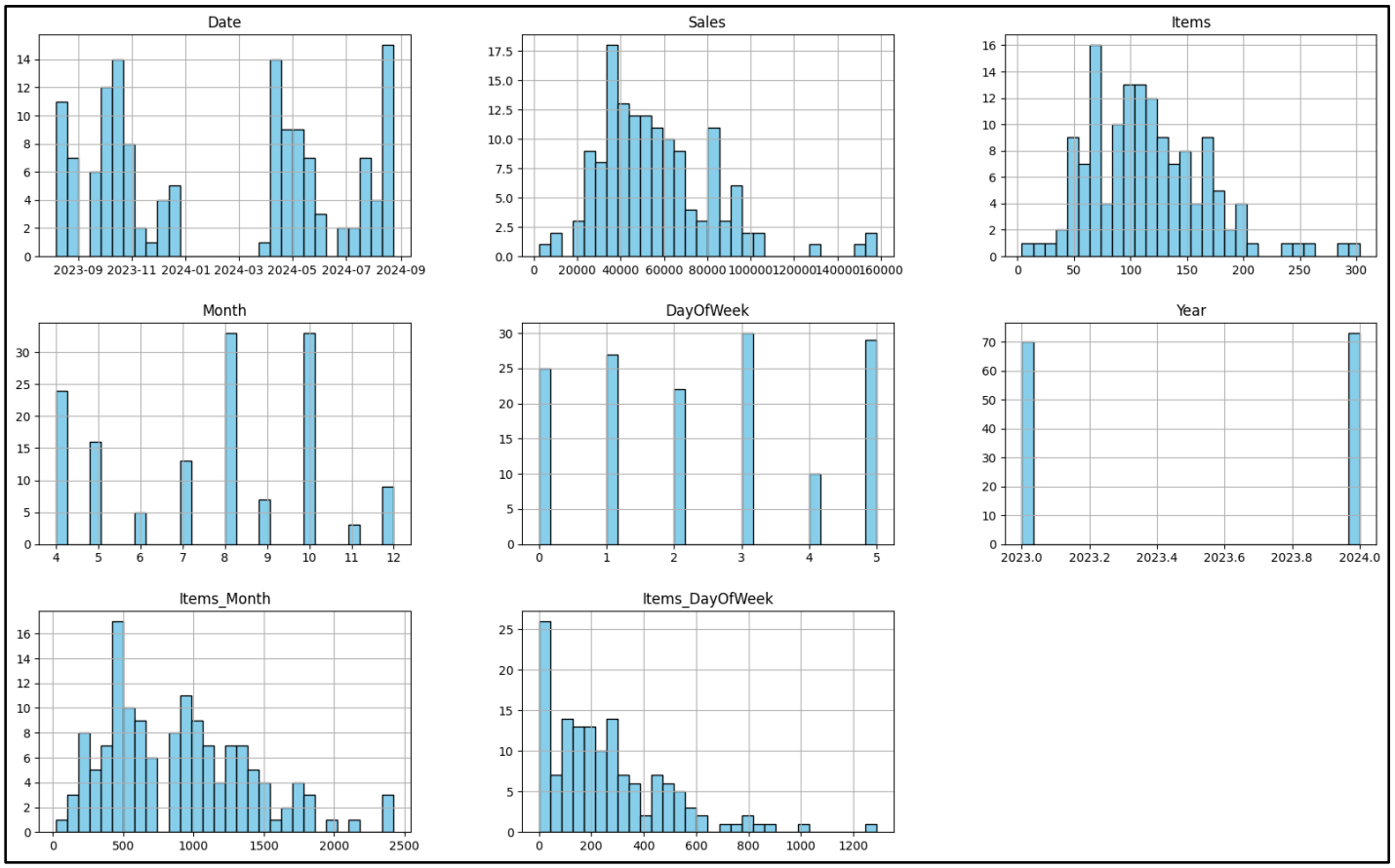
- The primary goal of this analysis was to check what companies gave the highest revenue to Nalanda Dresses and to visualise the geographical locations of all the companies.
- The dataset consisted of only the names of the cities/towns to which the warehouses of the companies belonged.
- The latitudes and the longitudes were manually input to perform a geo-analysis and create a heatmap of sales density across India to check the high sale areas across the country.
- Further, for Nalanda Dresses belonging to Pune, the distance from Pune to all the locations was calculated and then a hypothetical cost factor (10 rupees per kilometre by road) was introduced (as the original cost was unavailable) to see how much cost is incurred by the company to transport the goods via road to various locations.

Results and Findings:

Part 3.1: Demand Forecasting Based on Sales Prediction for Inventory Optimisation:

Data Preprocessing and Exploratory Data Analysis:

Before visualising the clean dataset, various steps were taken to extract more information from the existing features.



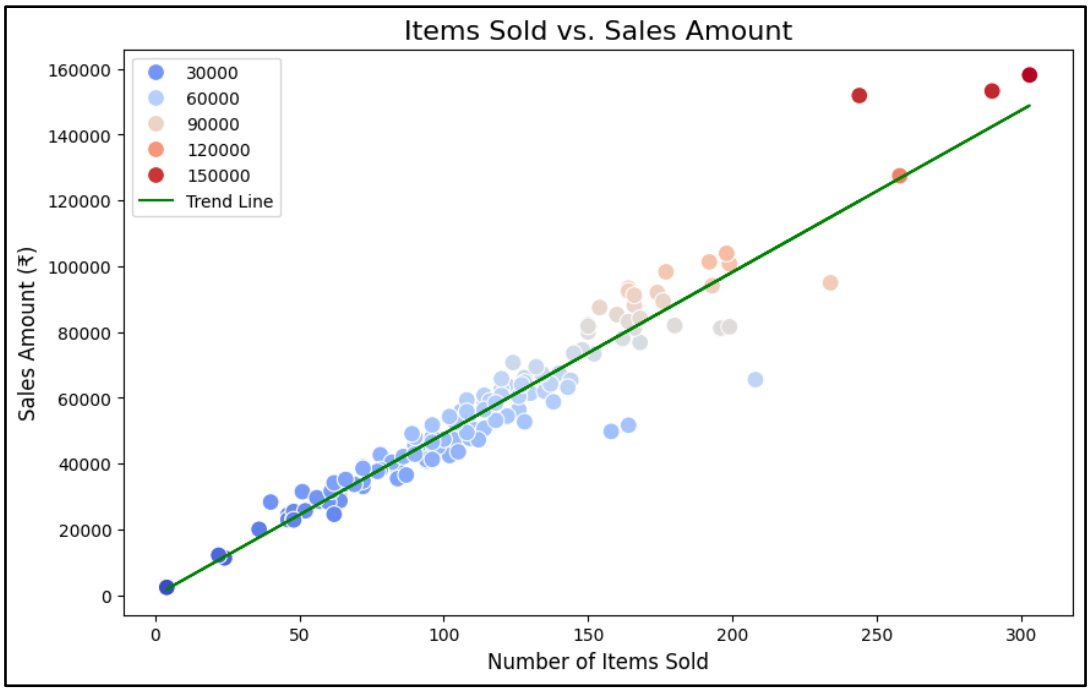
Graph 3.1: Histogram of all the numerical features in the dataset.

Graph 3.1 shows the features' data distribution and the target variable Sales. After observing the data distribution, it clearly shows that all the columns have different scales. So, for higher accuracy of the machine learning model, all the numerical columns are scaled using the Standard Scaler API.

Considering the columns of Items_Month, Items_DayOfWeek, and Sales, the data appears to be rightly skewed. Thus, a log transformation is applied to the skewed columns.

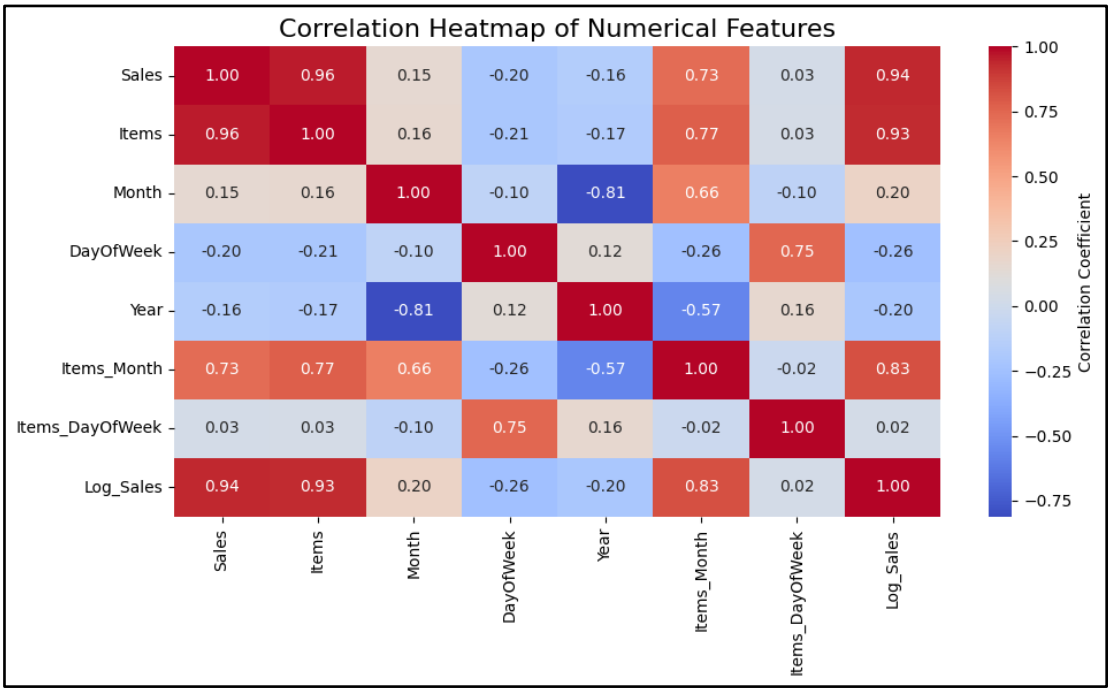
This graph is included in the final report to show the validity of further preprocessing applied to the dataset. Major decisions like log transformation and capping outliers are taken after looking at the scale and skewness of the data.

As mentioned before, the outliers in the dataset are capped using the winsorization technique. Winsorization replaces the specified limits with less extreme values. In this case, the top 1% and the bottom 1% of the data are winsorized. This helps maintain the structure of the data while reducing the influence of extreme outliers.



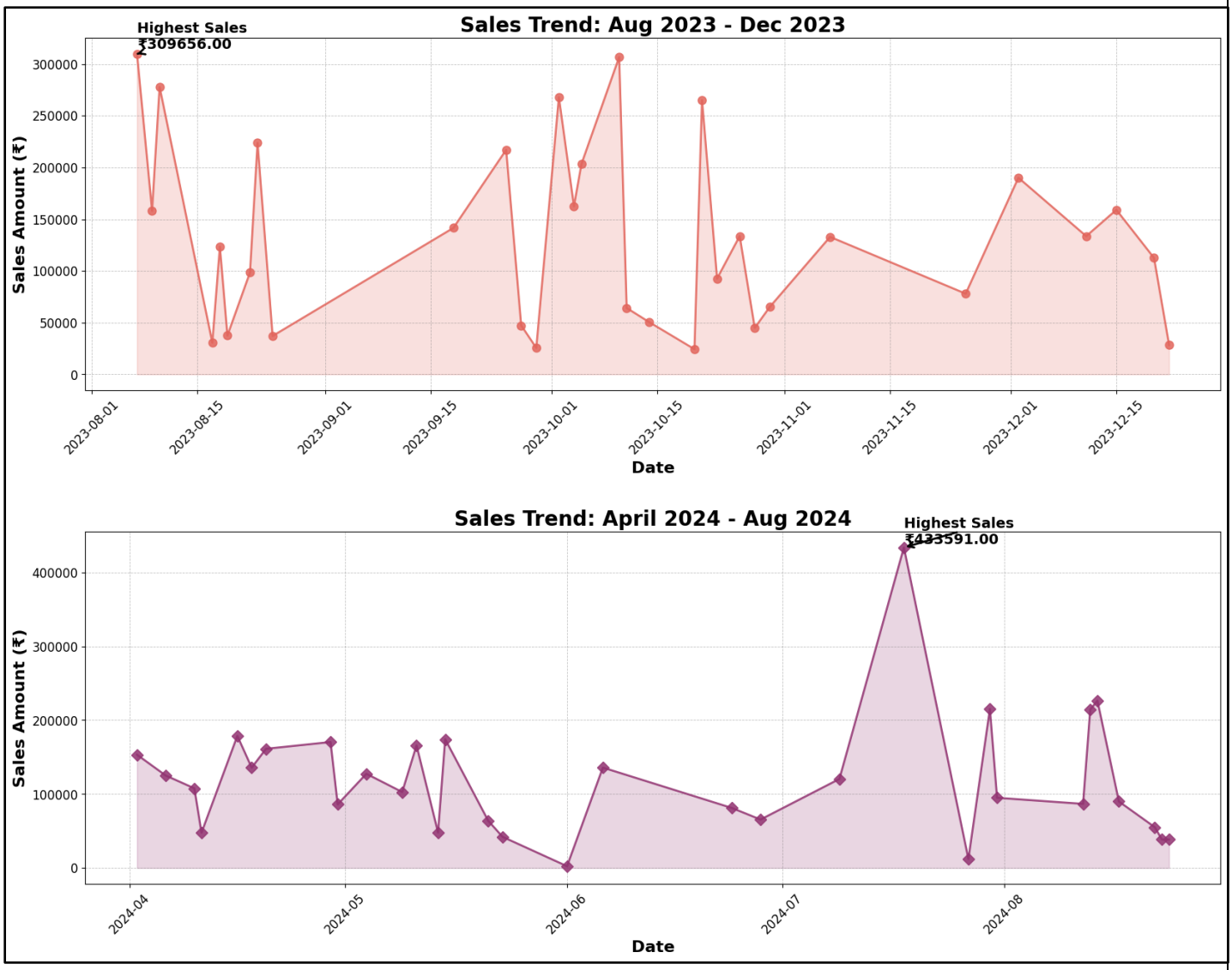
Graph 3.2: Scatter plot to examine the linear relation between Items and Sales

Graph 3.2 shows the linear relation between the target variable ‘Sales’ and the feature Items Sold. To show the linearity, a regression line is fitted on a scatter plot. The majority of clusters are around the regression line, which indicates a good fit.



Graph 3.3 Correlation Heatmap of all the numerical features

Graph 3.3 shows the correlation between all the numerical features (existing, extracted and interactive). Features like Sales, Items and Items_Month are strongly/positively correlated with each other exhibiting high feature importance. The negatively correlated features are kept as they may contribute towards model building.



Graph 3.4 Sales Trend Across Time

Graph 3.4 is a line plot that shows the sales of the company from the period of August 2023 – December 2023 and April 2024 – August 2024. The visible spikes are driven by the seasonal demand during the festivities like Ganesh Chaturthi and Diwali.

There is a larger spike in mid-July 2024 (₹4,33,591) compared to mid-September 2023 (₹3,09,656). The 40% increase in the highest sales value year-over-year suggests improved performance, likely due to stronger market strategies, increased demand, or promotions.

Model Building:

After applying the preprocessing techniques (scaling and encoding) and splitting the data into train and test datasets, a dictionary of all the major regression models is created to check the MSE (Mean Square Error) and the R2 scores across all the models.

Further analysis consists of two approaches, one without capping the outliers and the other with capping them.

Approach 1:

Model Name	Performance Metrics
Linear Regression	MSE: 0.015368373876855504 R2 Score: 0.971603872045381
Ridge	MSE: 0.03151400163289603 R2 Score: 0.9417716129305357
Lasso	MSE: 0.5606843497562463 R2 Score: -0.03597587262031832
Elastic Net	MSE: 0.5606843497562463 R2 Score: -0.03597587262031832
SVR	MSE: 0.33690446871328483 R2 Score: 0.37750197394190477
KNN	MSE: 0.2091703442618409 R2 Score: 0.6135161789032281
Decision Tree	MSE: 0.10581996276900543 R2 Score: 0.804476568111934
Random Forest	MSE: 0.15604339545557433 R2 Score: 0.7116787853201085
Gradient Boosting	MSE: 0.12959466472987805 R2 Score: 0.7605480767586249
XG Boost	MSE: 0.11040171583794789 R2 Score: 0.7960108678729456
Voting Regressor	MSE: 0.12517371173314118 R2 Score: 0.7687166668763962

Table 3.1: Performance Metrics without capping the outliers

Table 3.1 shows the MSE and the R2 score across all the regression models. The highest-performing models are Linear Regression and Ridge Regression.

The extremely low MSE and the R2 score of Linear Regression show that the accuracy of the model relies on the linear relationship between the features, and the model can explain 97% of the variance which suggests an excellent fit.

The second model that has the highest performance metrics is the Ridge Regression model. The initial structure of Ridge is similar to Linear Regression, except for a penalty (L2 regularization) that shrinks the coefficients towards zero without eliminating them, which helps in reducing the model's sensitivity to small changes in the data.

The objective function for Ridge Regression is:

$$\text{Minimize: } \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

Where λ is the Regularization parameter controlling the penalty strength. Ridge regression ensures robustness and the high accuracy achieved with the model shows that there is no overfitting.

Since Linear regression has no regularization parameter, a comparison between the train and test performance metrics is performed, to check whether the model overfits the data. This is done by comparing the MSE and the R2 scores of the train and test datasets. If the train MSE is less than the test MSE and the train R2 is more than the test R2 then the model appears to overfit. This is the result of the overfitting:

Training MSE: 0.008687571829655053

Training R2 Score: 0.9573866303393732

Test MSE: 0.015368373876855504

Test R2 Score: 0.971603872045381

The model does not appear to be overfitting.

This shows that the model has extremely high accuracy and performs well on unseen data.

Approach 2:

Model Name	Performance Metrics
Linear Regression	MSE: 0.0647100068131193 R2 Score: 0.7888998569375356
Ridge	MSE: 0.03582410330973331 R2 Score: 0.8831328614195727
Lasso	MSE: 0.31383920471659865 R2 Score: -0.023821573773227023
Elastic Net	MSE: 0.31383920471659865 R2 Score: -0.023821573773227023
SVR	MSE: 0.1026390388679794 R2 Score: 0.6651659170520825
KNN	MSE: 0.055965326445048695 R2 Score: 0.8174271801082228
Decision Tree	MSE: 0.036411608511814283 R2 Score: 0.8812162732701168
Random Forest	MSE: 0.022943592768350902 R2 Score: 0.9251522916733205
Gradient Boosting	MSE: 0.01566632000292336 R2 Score: 0.9488925661307615
XG Boost	MSE: 0.008565499273713476 R2 Score: 0.9720572101420988
Voting Regressor	MSE: 0.012983837709379003 R2 Score: 0.9576434908148698

Table 3.2: Performance metrics after capping the outliers

Table 3.2 shows the scores after capping the outliers using winsorization. All model performances have increased compared to the first approach and XG Boost, being a robust model that does not depend on linearity, gives the highest accuracy. XG Boost is an extremely powerful model that works on supervised machine-learning tasks.

Further improving the XG Boost model, hyperparameter tuning is done to increase the accuracy of the model.

The best parameters found:

{'alpha': 0, 'colsample_bytree': 0.8, 'lambda': 2, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.7}.

Alpha being 0 suggests that the model performs better without regularization.

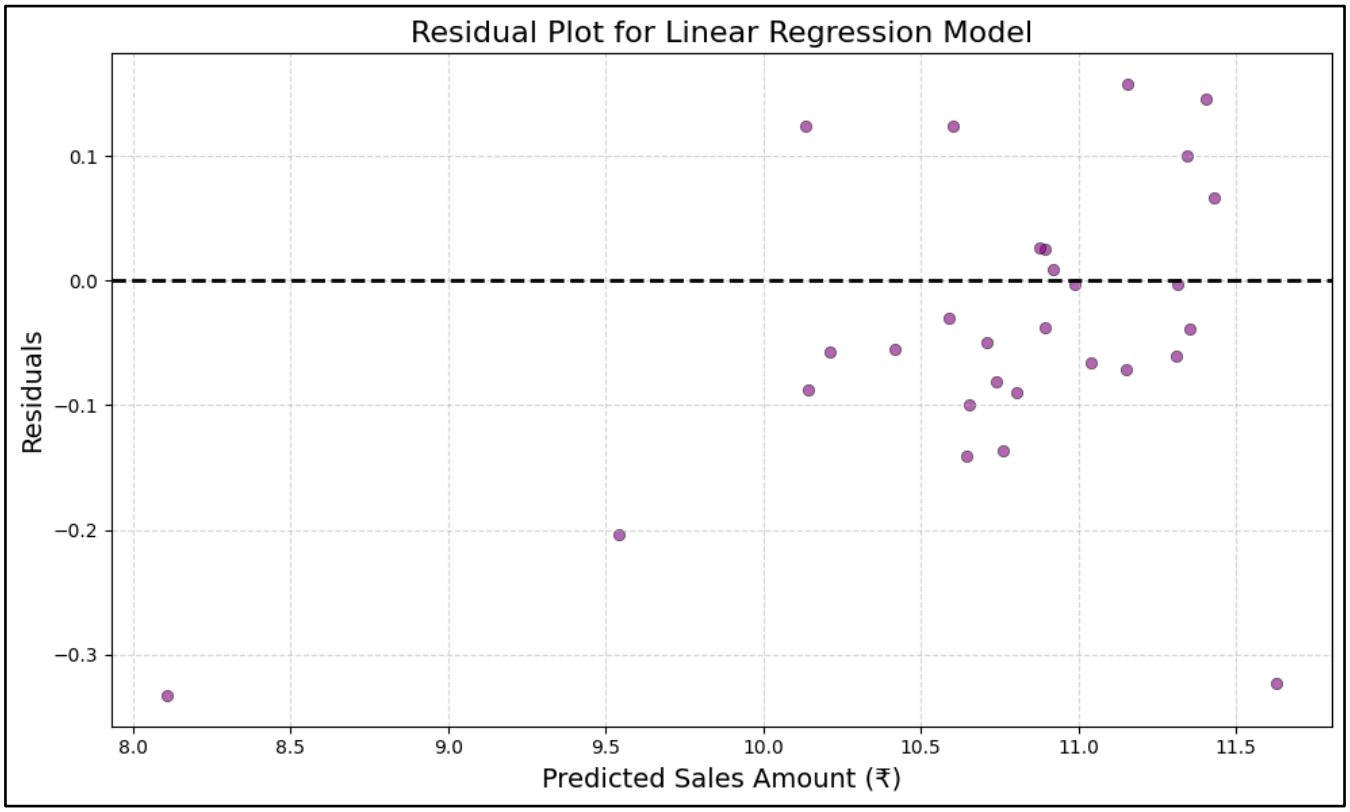
Training MSE: 0.002144087427356198, Training R2 Score: 0.9894516139233007.

Test MSE: 0.010447737673983713, Test R2 Score: 0.9659168801507542.

The training score is slightly better than the test score. This is because tree-based models tend to overfit. Looking at the training and testing values the difference is negligible. Thus, XG Boost proves to be a good fit for predicting sales.

Model Evaluation:

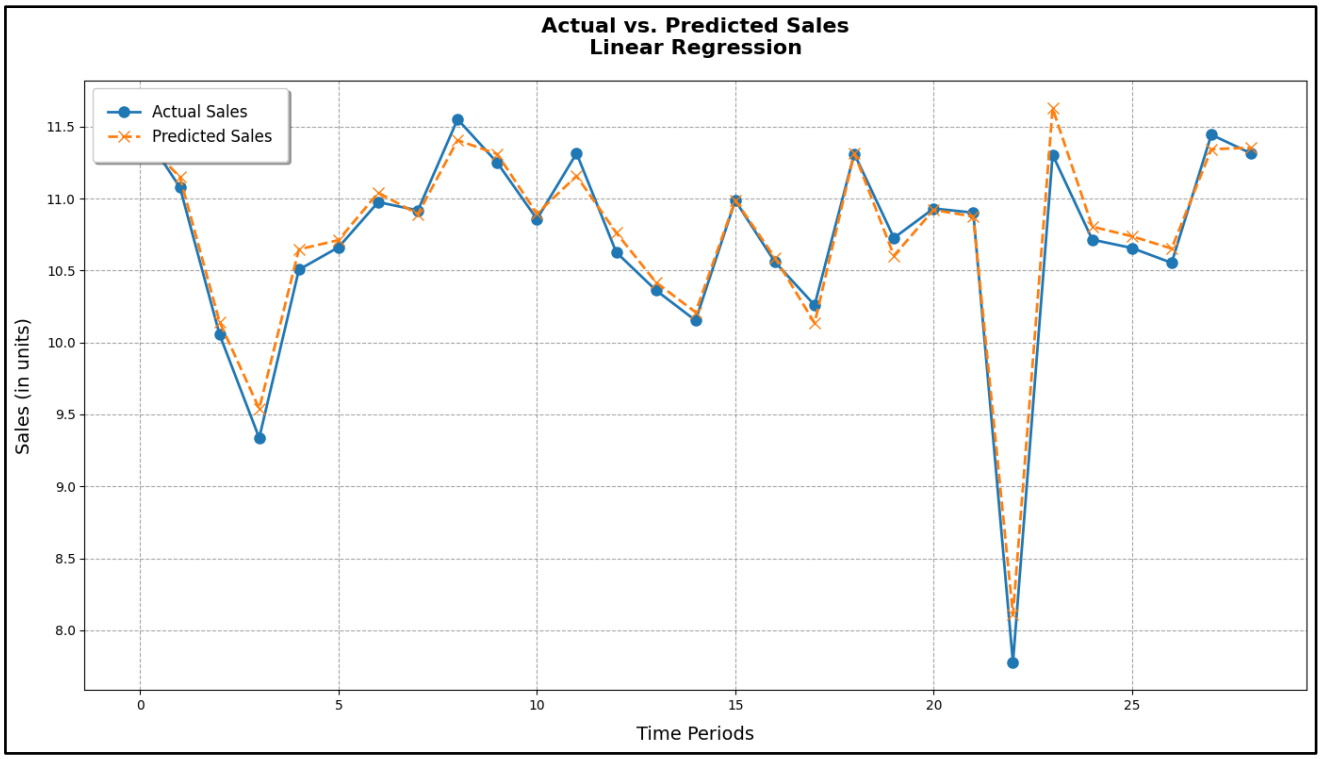
The final step for the Sales Prediction is to check how accurately the model performs. This accuracy is determined by plots like Residual Analysis, Feature Importance, and lastly plotting the time series data of the actual vs the predicted sales.



Graph 3.5 Residual plot for Linear Regression (Approach 1)

After predicting the target variable, the difference between the actual values and the predicted values is calculated by the residual plot in 3.5. The residuals (data points) are randomly scattered near y=0. This indicates that it is a good fit since there is no visible pattern.

There are a few extreme outliers near points 8.0 and 11.5 which shows the model may struggle with extreme values. Overall, the model can predict the sales accurately with little no to difference.

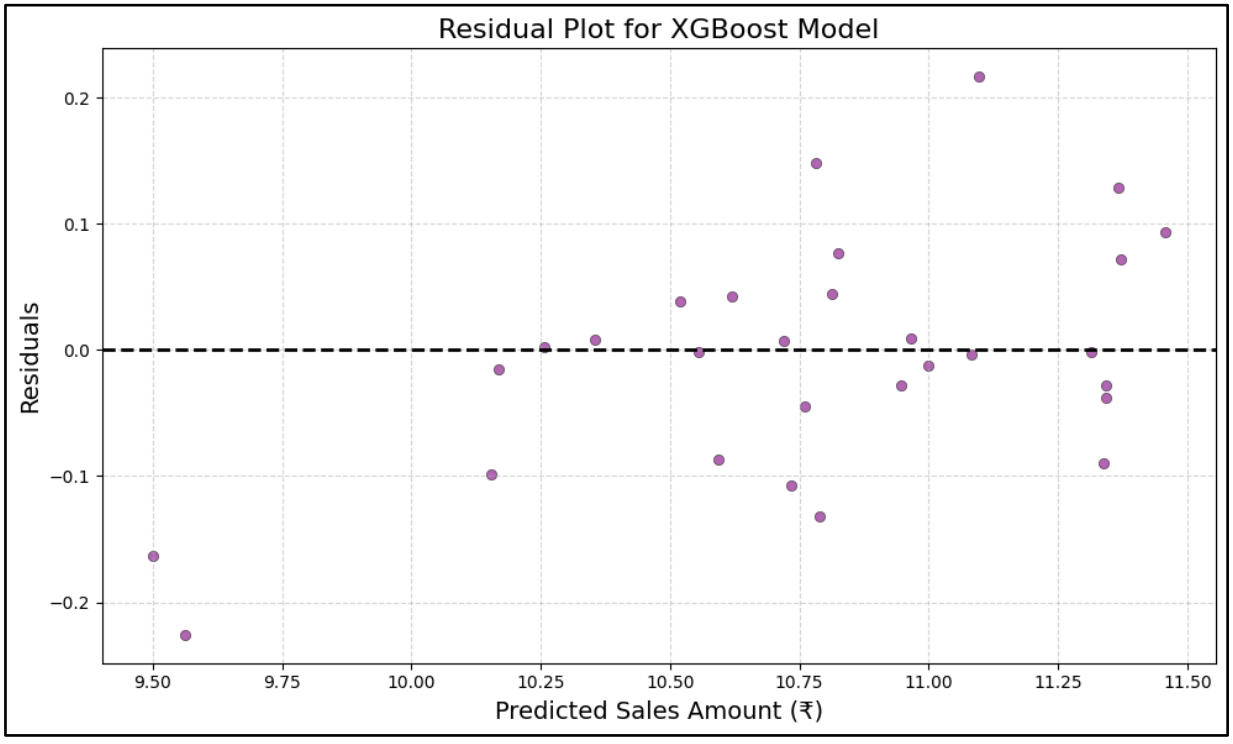


Graph 3.6: Actual Vs Predicted Values (Sales Prediction)

Graph 3.6 is a line plot that shows the difference between the actual sales values (after transformation) and the predicted values by the Linear regression model.

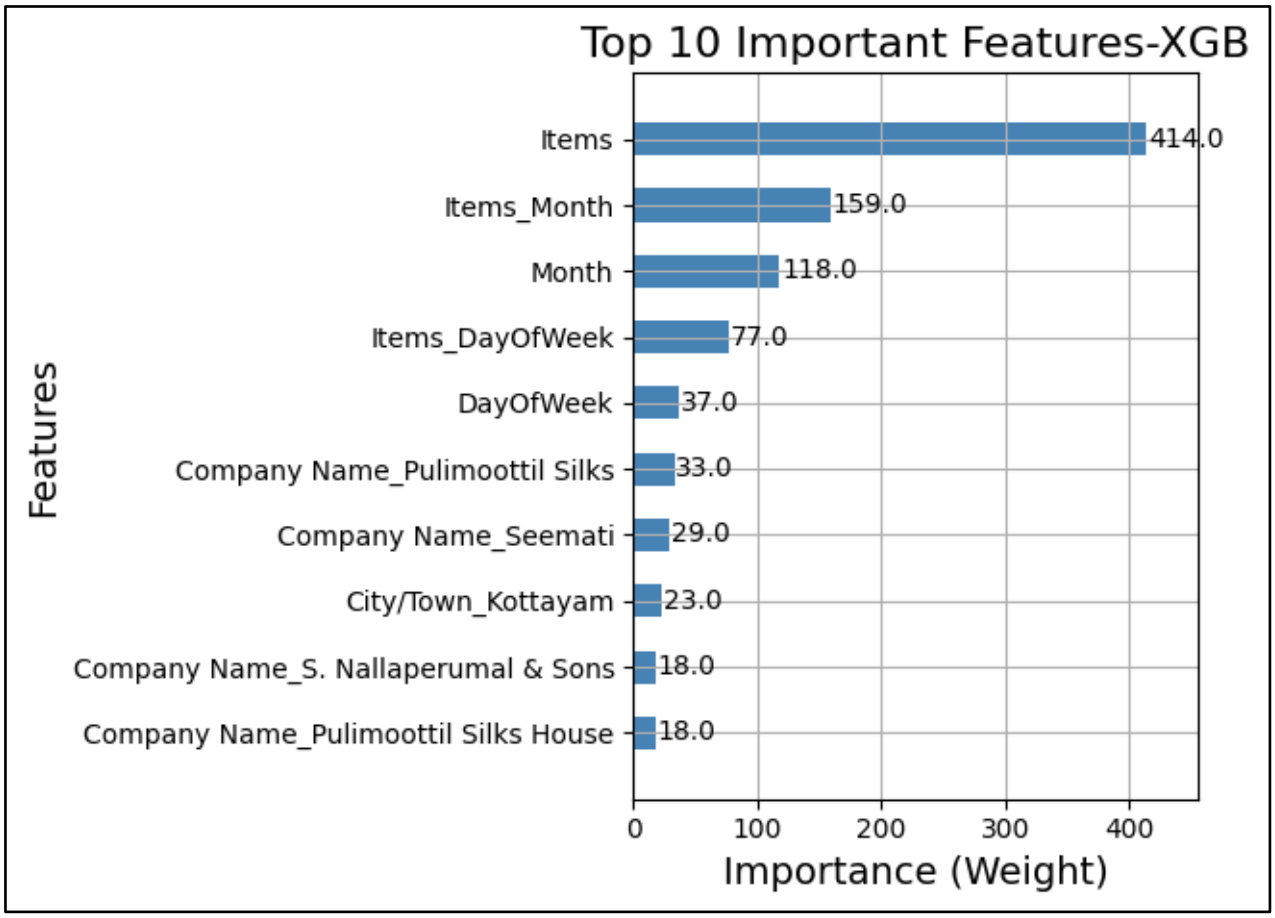
For most of the time period, the predicted values are extremely close to the actual values on the plot. This graph is the ultimate showcase of how well the model can predict the sales figures.

Between the time periods 1–10 and 15–20, the predicted and actual values align almost perfectly. Although the model is not able to capture the steep changes in sales, the Linear Regression model performs well for the majority of the time periods, indicating that the data has a strong linear trend that the model captures effectively.



Graph 3.7 Residual Plot for XG Boost (Approach 2)

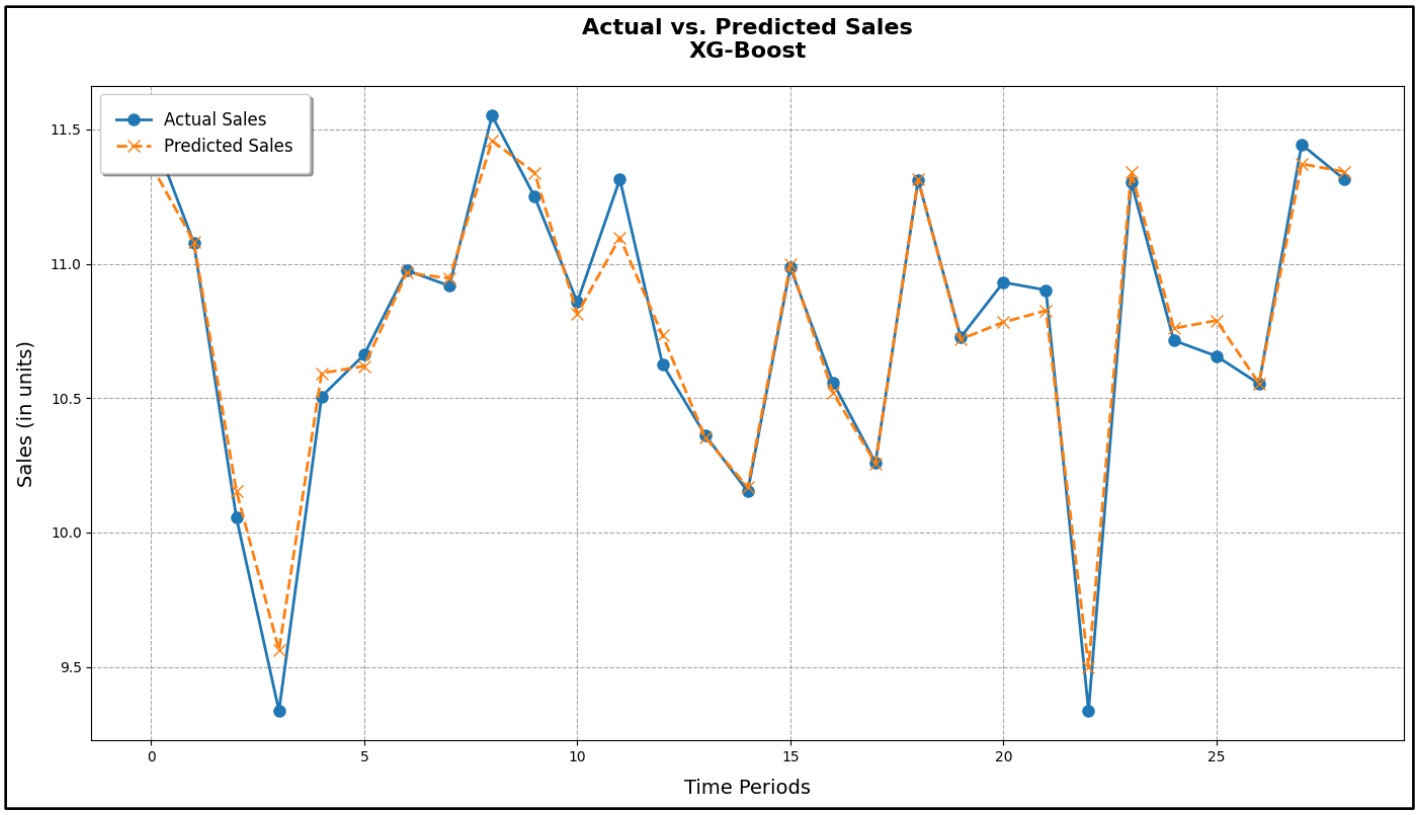
The second residual plot is created by calculating the difference between the true and the predicted values using XG-Boost. There is a clear difference between the two plots. In Graph 3.7 the residuals are more tightly clustered to y = 0 and show no clear pattern, which is an indication of a good fit. Again, similar to Linear Regression there are a few outlier residuals near 9.50 and 11.0, whose residual values go beyond the +-0.2. This is an indication that the model might be slightly sensitive to extreme values. Overall, the residual plot exhibits an excellent fit, with a difference of around +-0.2 between the actual and predicted values.



Graph 3.8 Feature Importance for the XG Boost Model.

The feature importance of the XG Boost model is shown in Graph 3.8. Due to a strong correlation with the target variable (Sales), the feature 'Items' has the most significance (weight = 414). The other important features include Items_Month, Month, Items_DayOfWeek and Day_OfWeek, this suggests a seasonal implication for sales.

Other important features include companies like Pulimoottil and Seemati, which is due to the company's performance, and locations like Kottayam that heavily influence sales since the majority of the warehouses belong to these locations.



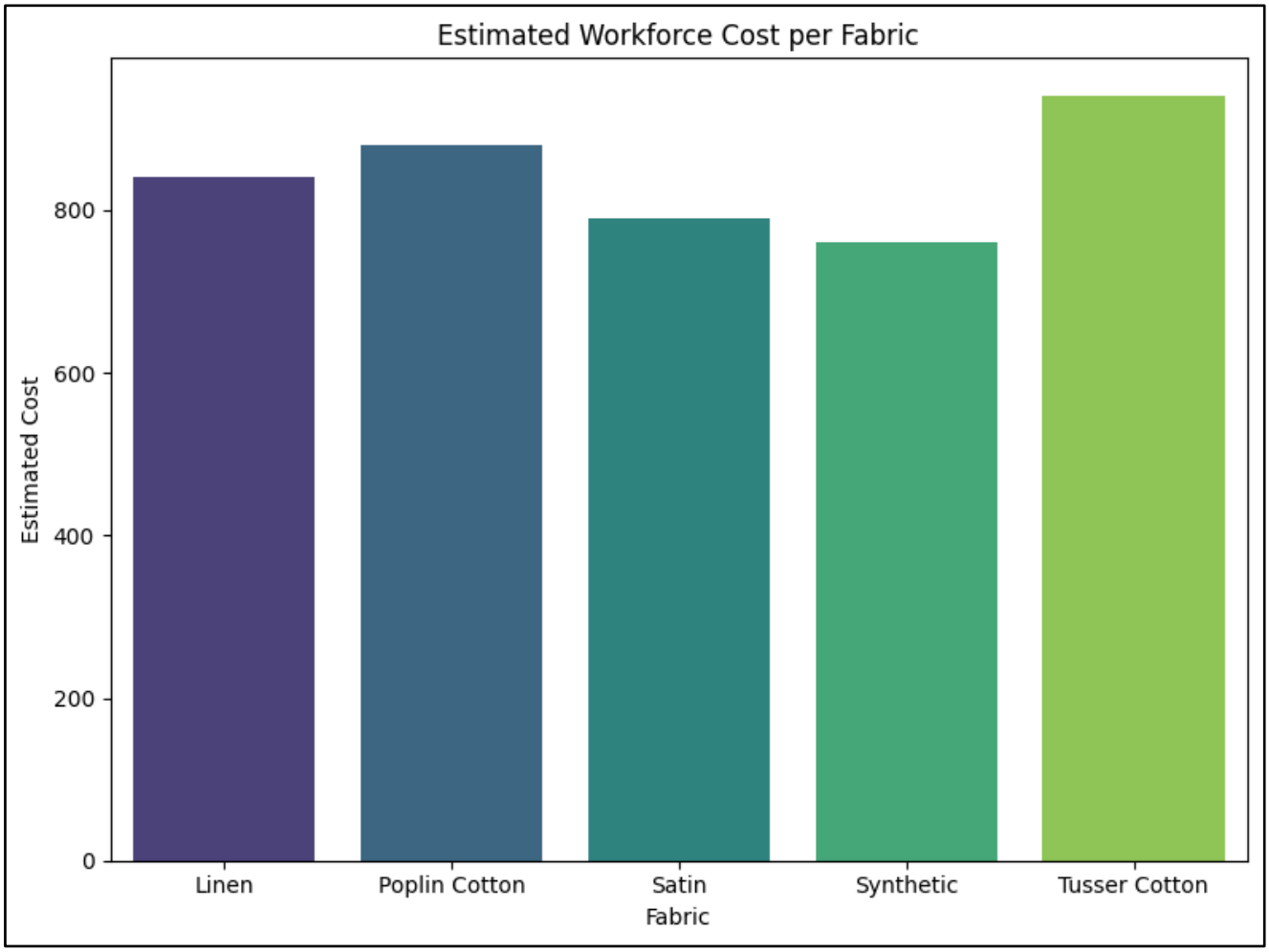
Graph 3.9 Final Actual vs Predicted Values (Sales Prediction)

The predicted values in Graph 3.9 align almost perfectly, especially when there aren't any sudden climbs or drops in sales units, with the true values. There are a few miscalculations, like points 20 and 25. However, the XGB model is more successful in capturing the actual data points during sudden peaks and troughs compared to Linear Regression.

	Actual Value	Predicted Value
117	11.495586	11.367411
19	11.078521	11.082476
82	10.055608	10.153951
97	9.336180	9.562231
56	10.505971	10.592809
12	10.661814	10.619226
131	10.975465	10.966566
65	10.917758	10.945668
66	11.551069	11.457475
18	11.249350	11.338659
51	10.857325	10.813254
78	11.314804	11.097923
94	10.626533	10.733918
132	10.362778	10.354819
100	10.152572	10.168365
64	10.986191	10.998542
27	10.558907	10.519955
69	10.259377	10.257019
125	11.311727	11.313100
73	10.726566	10.719192
11	10.930998	10.782425
119	10.902022	10.825690
110	9.336180	9.499882
113	11.304510	11.342051
55	10.714507	10.759567
45	10.656388	10.788762
9	10.553336	10.554924
4	11.443843	11.371713
26	11.314340	11.342871

Image 3.1: A Table consisting of the actual sales units vs predicted sales units

Part 3.2: Workforce Cost Analysis:



Graph 3.10 Bar chart showing the estimated workforce cost for each dress material.

The most cost-heavy fabric is tusser cotton, while synthetic is the least. The Workforce Cost estimate appears to be uniform across all the fabrics. The data-entry Synthetic Jacket hasn't been included in the final plot as it acts as a false outlier.

Part 3.3: Analysing the Geographical Locations and the cost incurred for transport.

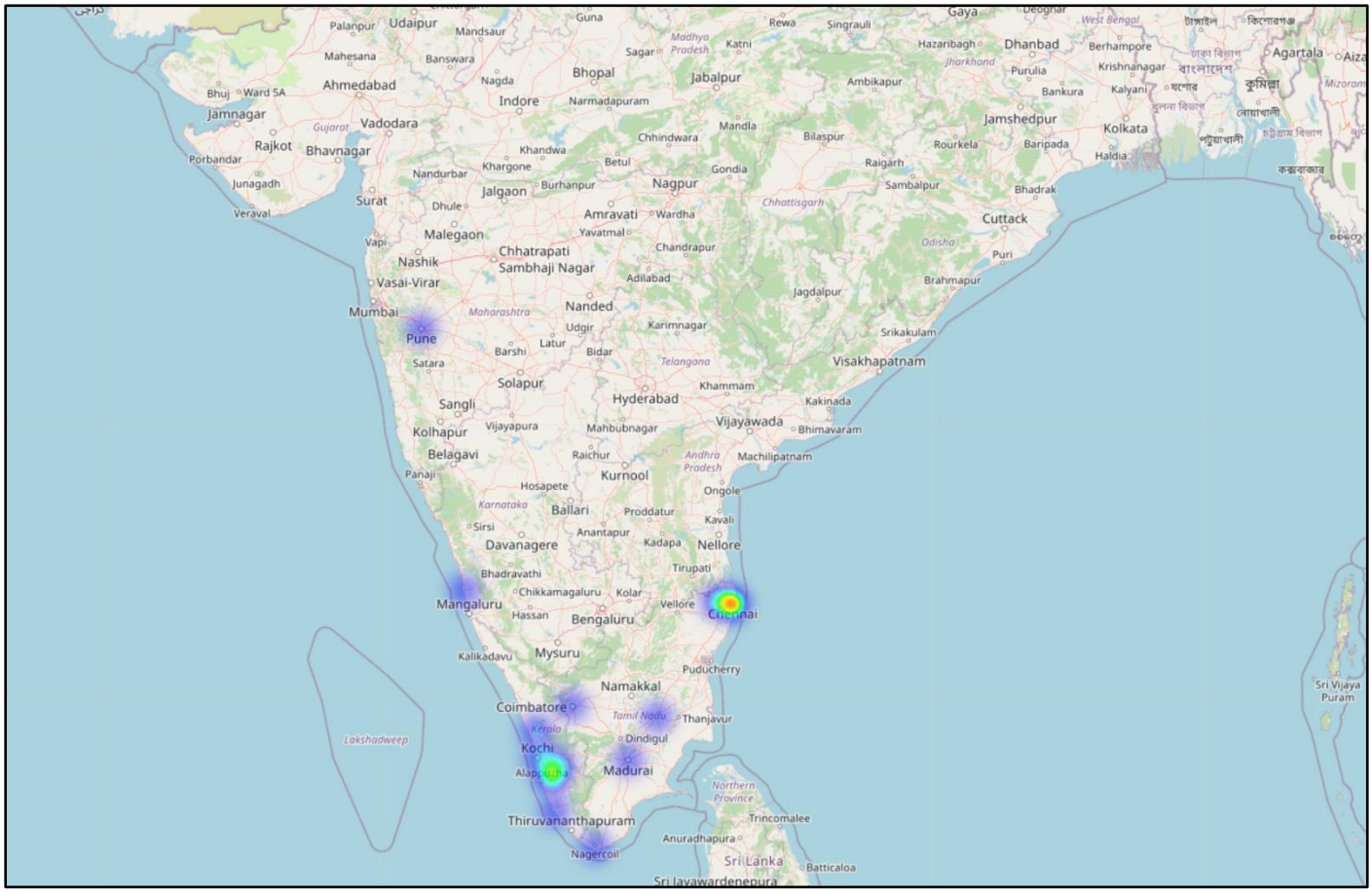


Image 3.2: A screenshot of the heatmap created to highlight the sales density across India

Click here to see the web view created of Image 3.2. As one can observe through the heatmap, the majority of sales occur in the southern part of India.

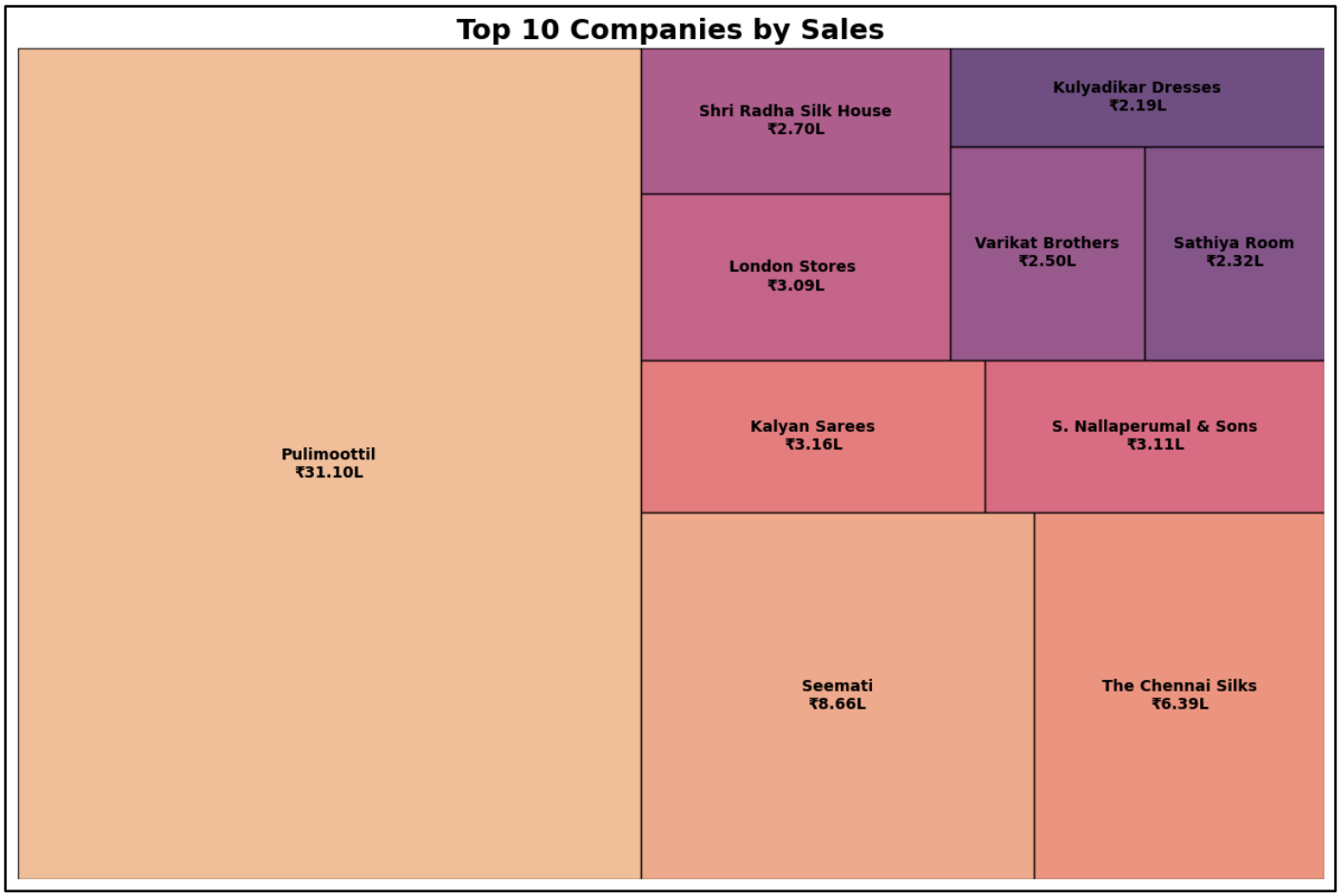
Chennai and Kottayam are the cities having the highest sales density. Cities like Nagercoil, Tiruvallur, and Thrissur have moderate sales density. And the western part of India (where the home company is located), one can see negligible to no sales.

The estimated cost incurred by the company in transporting goods to these locations is shown in this: Transport Cost.

Apart from the web view, this table shows the estimated cost incurred:

City/Town: Chengannur	Distance: 1061.06 km	Transport Cost: ₹10610.56
City/Town: Chennai	Distance: 913.14 km	Transport Cost: ₹9131.41
City/Town: Chrompet	Distance: 912.98 km	Transport Cost: ₹9129.76
City/Town: Coimbatore	Distance: 894.69 km	Transport Cost: ₹8946.88
City/Town: Ernakulam	Distance: 980.78 km	Transport Cost: ₹9807.84
City/Town: Kollam	Distance: 1106.0 km	Transport Cost: ₹11060.05
City/Town: Kottayam	Distance: 1028.9 km	Transport Cost: ₹10288.96
City/Town: Madurai	Distance: 1056.23 km	Transport Cost: ₹10562.33
City/Town: Nagercoil	Distance: 1207.75 km	Transport Cost: ₹12077.54
City/Town: Pala	Distance: 1021.8 km	Transport Cost: ₹10217.97
City/Town: Pune	Distance: 0.0 km	Transport Cost: ₹0.0
City/Town: Thodupuzha	Distance: 1003.23 km	Transport Cost: ₹10032.35
City/Town: Thrissur	Distance: 920.09 km	Transport Cost: ₹9200.88
City/Town: Tiruvallur	Distance: 879.7 km	Transport Cost: ₹8797.05

City/Town: Trichy	Distance: 1001.87 km	Transport Cost: ₹10018.67
City/Town: Udupi	Distance: 580.94 km	Transport Cost: ₹5809.37
City/Town: Velacherry	Distance: 917.42 km	Transport Cost: ₹9174.23



Graph 3.11: Tree Map showing the top companies contributing to the sales of the company.

Note: Pulimoottil data is combined as the customer company has various warehouses across the country.

After looking at the tree map in 3.11, the largest sales contributor to Nalanda dresses is Pulimootil with 31.10 lakhs of sales with approximately 40 percent of sales.

The treemap gives a clear distinction between customer companies making it easy to identify the top contributing companies sales-wise.

Interpretations and Recommendations

1. The majority of the analysis was conducted by creating a sales prediction machine-learning model that would be able to predict sales based on features like items (order placed), seasonal fluctuations, etc.
   After creating two ML models with two approaches, the models can predict sales with up to 97 per cent accuracy on unseen data.
   Both the ML models (Linear Regression and XG-Boost) have extremely low Mean Square Error and high R2 score.
   This allows a data-driven approach to inventory management, which was the primary problem of the organisation as they relied on a demand-and-supply approach.
   The final visualizations included 2 major plots:

   - Feature Importance Plot: Items, the interactive features between items and date, and the date were the most important features that contributed to the model building part. This result shows that sales are heavily influenced by items, which is an obvious fact due to the linear relation between the two variables, as well as the seasonal trends.
   - Sales Prediction Line Plot: The line plot was able to predict the sales units with extreme accuracy, with a slight deviation during fluctuations (positive and negative peaks).

Based on the results acquired by examining the performance of the machine-learning model. The business should prioritize stocking up after analysing the trends of the scalable machine learning model. If the business can create new features that segregate the items based on their cost price the model would further benefit from the information.

The months and the items_month being important features suggests that the sales are impacted due to festivity and seasonal trends. The business should identify the peak season for production and restock on inventory or manage inventory levels based on the trends. This will reduce the workload of the labourers and prevent loss of labour.

Lastly, during high sales season, the business could consider outsourcing the production to other local organisations to reduce the workload amongst the existing employees.

2. The second major result of the analysis conducted was the geo-analysis of the customers of the company. After observing the heatmaps that show sales density across different cities and towns in India, the result shows that the company has around 95 per cent of the customers belonging to the southern part of India.
   The most sales-dense cities include Chennai and Kottayam, and the rest are scattered throughout the southern states.
   There was only one customer company called Sai Baba stores which the company supplied locally.
   A hypothetical factor was created that determines the cost incurred by the company to transport the goods to the warehouses of the customer companies, this created data gave a scenario of how much the company has to suffer in monetary value to transport the goods. For example, the cost incurred for transporting the goods to Chennai is 9,131.4 rupees.

The company can adopt a marketing strategy to acquire customers locally and build a presence in the State of Maharashtra.

This will reduce the travel costs for marketing as well as delivering the products to distant companies that belong to the southern parts of India.

Though distant the frequency of the same customer companies is present throughout the period of 10 months. This suggests that the company has a loyal customer base and won't lose its current customers, so the company should market their products locally to acquire new customers.

Proof for the Analysis Conducted:

Part 1:

https://colab.research.google.com/drive/16JbQmFK-MXH0KgBMZ1S0KeEvGarBFVk2?usp=sharing

Part 2 and 3:

https://colab.research.google.com/drive/1lKHLZXrxxuyL67OqSHNtMPKMloh2tosU?usp=sharing