**Problem Statement**

**Business Report**

**Supply Chain Management**

****

**Problem Statement-A Fast Moving Consumer Goods (FMCG) company entered into the instant noodles business two years back. Their higher management has noticed that there is a mismatch in the demand and supply. Where the demand is high, supply is pretty low and vice-versa which results in a loss in inventory cost and ultimately loss to the company. Hence, the higher management wants to optimize the supply quantity in each and every warehouse in the entire country.**

**GOALS-**The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time from the respective warehouse.

1. Focus on all steps of data science (EDA, data processing, model, evaluation, charts)
2. Highlight any trend in data, deep insight, novel steps that you take
3. Highlight next steps and improvements.
4. Apply 5 to 6 machine learning algorithms and evaluate it

The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time from the respective warehouse.

### **ABOUT DATASET-** Dataset is a CSV file called Data-1.csv This file contains 250000 rows and 24 attributes. The dataset for this problem consists of various features such as warehouse ID, manager ID, zone, regional zone, number of refill requests received by the warehouse in the last 3 months, number of transport issues for the warehouse in the last 1 year, number of competitors in the market, number of retail shops who sell noodles produced by the warehouse, whether the warehouse is owned by the company or it is on rent, number of distributors who work between warehouse and retail shops,whether the warehouse is in a flood impacted area or not, whether the warehouse has proper electric supply along with some power backup, distance from the warehouse to production hub, number of workers in the warehouse, warehouse establishment year, storage issues reported by the warehouse in the last 3 months, whether the warehouse has temperature regulating machine indicator or not,type of approval warehouse having been issued by government, number of times the warehouse faces the breakdown in the last 3 months, product weight, etc.

**1.Project Overview**

2.Data Exploration

3.Data Cleaning

4.Feature Engineering

5.Exploratory Data Analysis

6.Data Preprocessing

7.Model Building & Evaluation

8.Evaluation

9.Conclusion This project focuses on improving supply chain management using supervised regression analysis. It involves the following key steps:

Data preprocessing and cleaning to handle missing values and irrelevant features.Feature engineering to create new variables and enhance the dataset. Exploratory Data Analysis (EDA) to gain insights and understand the relationships between different features. Model building using various regression techniques such as Linear Regression, Decision Tree, Random Forest, AdaBoost, and Gradient Boosting. Evaluation of the regression models to determine their effectiveness in predicting and improving supply chain efficiency.

## Data Exploration

### Data Dictionary-

* Ware\_house\_ID : Unique Warehouse id where product is prepared for dispatch
* WH\_Manager\_ID : Manager Id present in the warehouse
* zone : Zone of the Warehouse
* WH\_regional\_zone : Regional Zone of the warehouse
* num\_refill\_req\_l3m : Refilling request received by the warehouse in the last 3 months
* transport\_issue\_l1y : No. of transport issued for warehouse in last 1 year
* Competitor\_in\_mkt : No. of competitors in the market
* retail\_shop\_num : Number of retail shops who sell noodlesproduced by the warehouse
* wh\_owner\_type : The warehouse is owned by the company or it is on rent
* distributor\_num : No. of distributor who works between warehouse and retail shops
* flood\_impacted : Is the warehouse in a flood impacted area or not
* flood\_proof : Warehouse is having flood proof indicator
* electric\_supply : Does the warehouse have proper electric supply along with some power backup
* dist\_from\_hub : distance from the warehouse to production hub
* workers\_num : no. workers in the warehouse
* wh\_est\_year : warehouse establishment year
* storage\_issue\_reported\_l3m : storage issues reported by the warehouse in the last 3 months
* govt\_check\_l3m : Government checking in last 3 months
* temp\_reg\_mach : warehouse having temperature regulating machine indicator or not
* approved\_wh\_govt\_certificate : Type of approval warehouse having been issued by government
* wh\_breakdown\_l3m : Number of times the warehouse faces thebreakdown in the last 3 months
* product\_wg\_ton : Product weight
* Dataset has 25 columns and 16620 rows.
* No. of columns with int datatype : 15
* No. of columns with object datatype : 8
* No. of columns with float datatype : 2

### **APPROACH**

**1.UNDERSTANDING DATA**- Python programming language will be used to conduct both visual and non-visual understanding of the data

**2.DATA PREPARATION**- In this step we will prepare the data for modelling. The cleaning, Pre-processing of data will happen in this stage.

**3.MODELLING**- Different regression models and tuning techniques are used until best fit model for our problem is obtained.

**4.EVALUATION**- At this step we will broadly look at all our models to find out which model meets the requirement of our business problem

**5.DEPLOYMENT**- Last step would be to document our project.

## Data Preprocessing

This project mainly focus on analyzing impact of features provided in the dataset on target variable product weight.

### 1. Drop irrelevent columns

1. Ware\_house\_ID
2. WH\_Manager\_ID

These three columns each contain entirely distinct values. In the context of this project, the uniqueness of these columns holds no relevance, as they are only analyzing product quantity. Therefore, these three columns have been removed.

### 2. Missing values

Among the 25 columns, null values are present in three columns. These columns correspond to the number of workers in the warehouse (workers\_num), the year of warehouse establishment (wh\_est\_year), and the grade of government certificate approved for the warehouse (approved\_wh\_govt\_certificate).

% of null values in eachg column:

workers\_num(4.01 %) wh\_est\_year(47.29 %) approved\_wh\_govt\_certificate(3.6 %)

**workers\_num :**

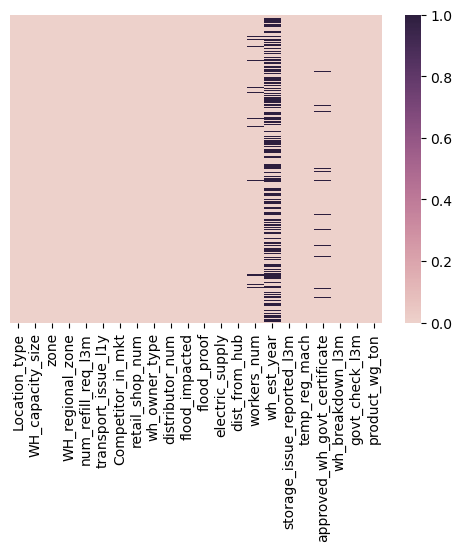
* numerical column
* positively skewed Filled null values with median.

**wh\_est\_year :**

Since half of the values in the column wh\_est\_year are null, decided to drop it.

**approved\_wh\_govt\_certificate:**

* categorical column Filled null values with mode

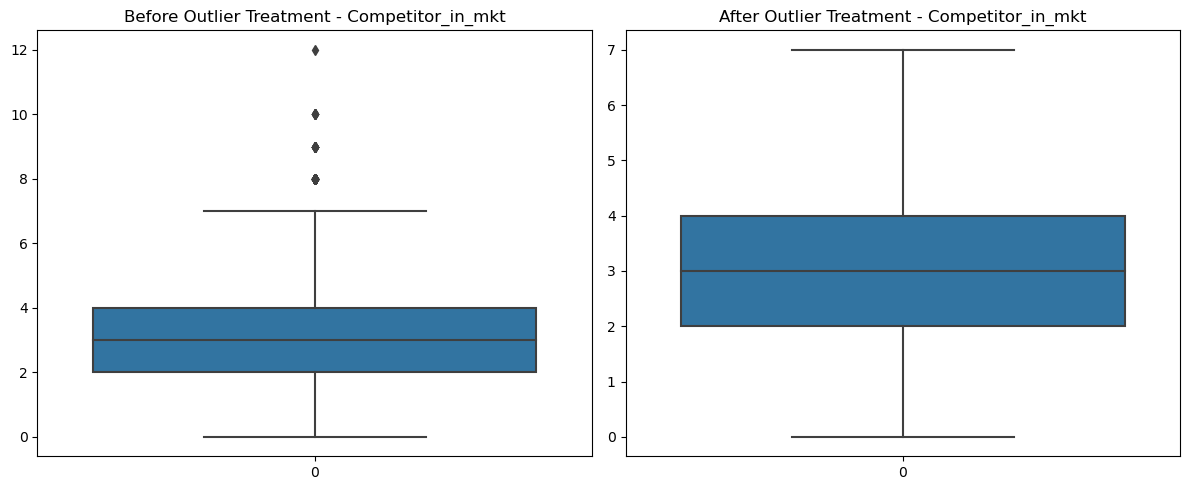


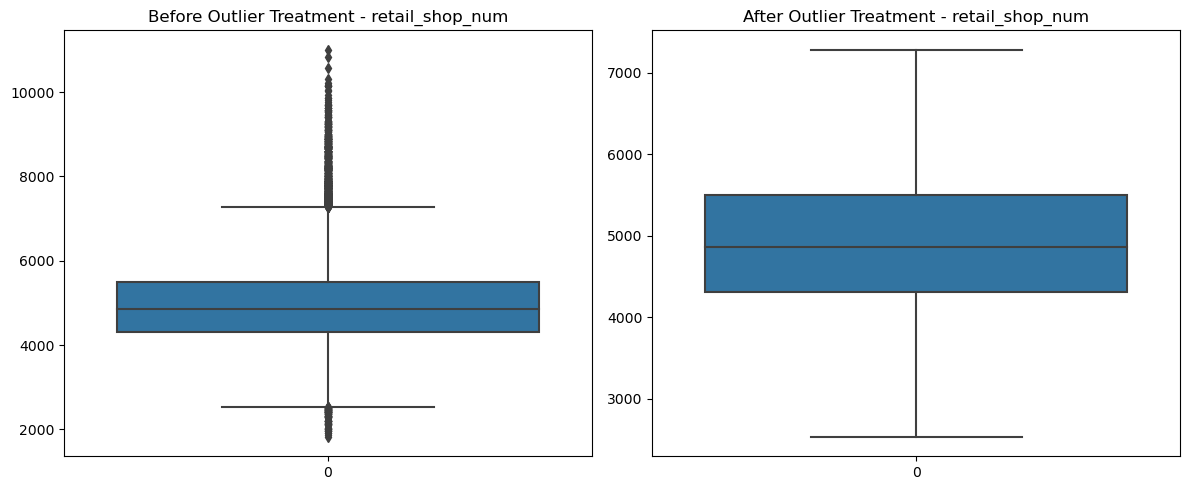
### 3. Duplicate values

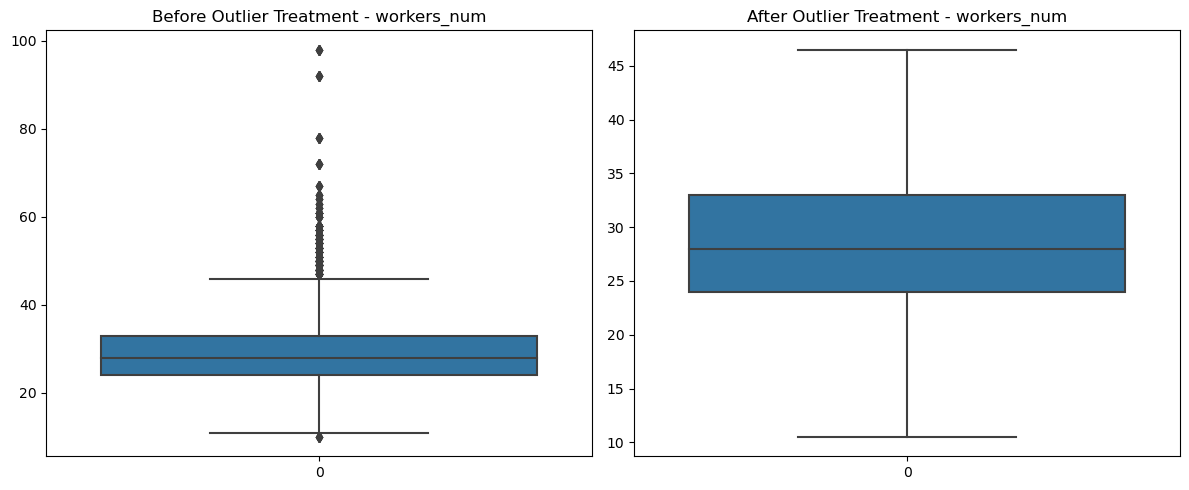
No dupliocate values were present in the dataset.

### 4. Outliers

Outliers were present in some of the columns. plotted boxplot to locate outliers.





Here we determined the count of outliers in every column. Among them, the 'flood\_proof' and 'flood\_impacted' columns, each containing only two distinct values, exhibit significant disparities in their value distributions and lack correlation. one of the unique values is considered as outlier due to its significantly lower proportion. Consequently, both of these columns were removed from the dataset.

**Outliers were detected and removed**

**Data Exploration** Exploring the dataset to understand the characteristics of the data is a crucial step in the project. This includes examining data statistics, visualizing key features, and identifying correlations.

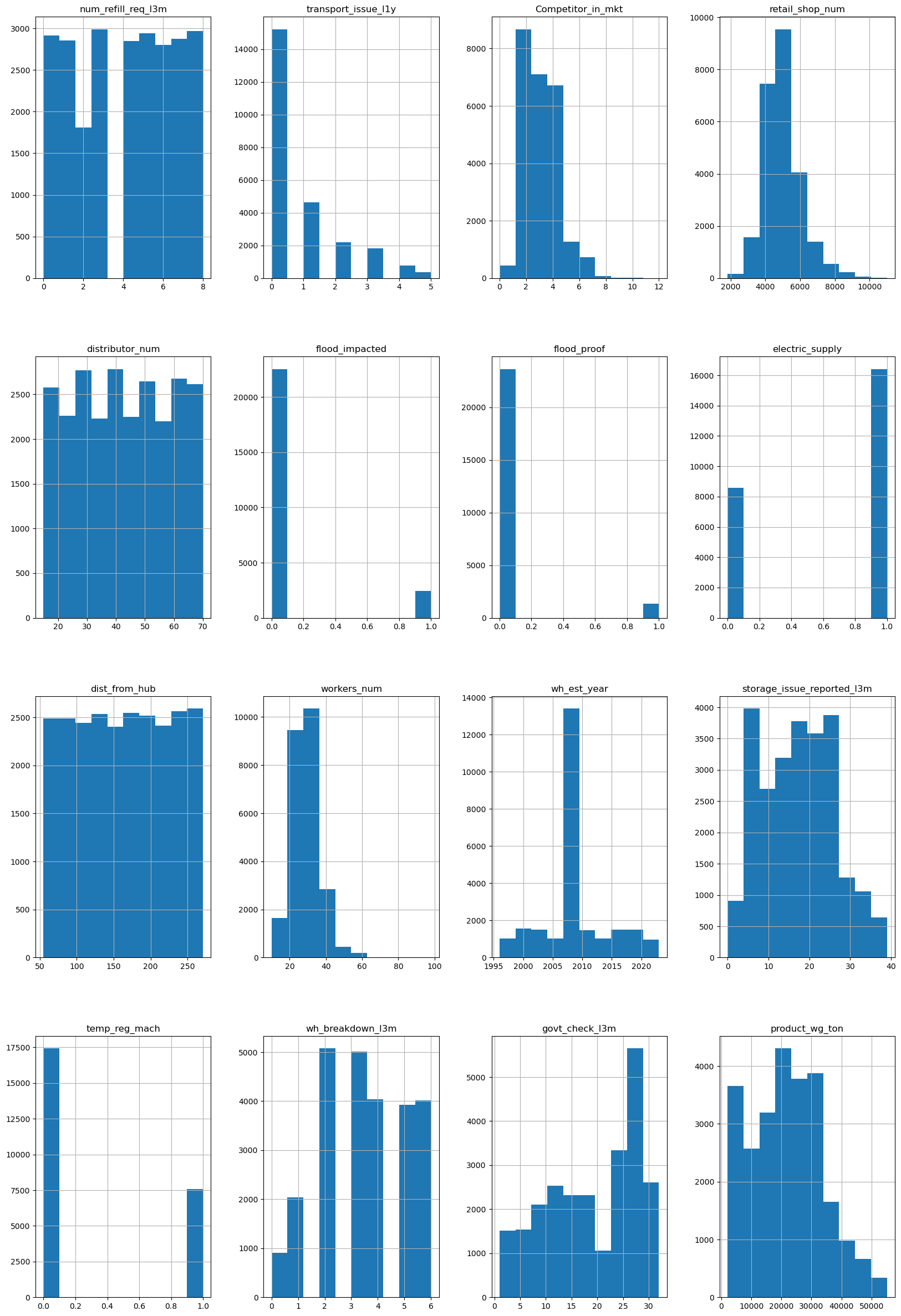
**Data Cleaning** Data cleaning is essential to ensure that the dataset is free from missing values and irrelevant information. This step involves handling null values, removing unnecessary columns, and preparing the data for analysis.

**Feature Engineering** Feature engineering involves creating new variables or transforming existing ones to enhance the dataset's predictive power. In this project, features like warehouse age, demand, supply, and demand interaction have been created to better understand the supply chain dynamics.

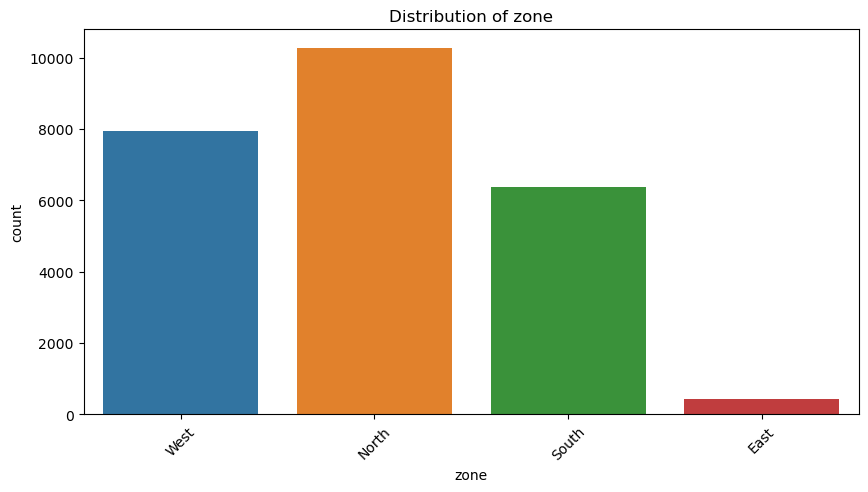
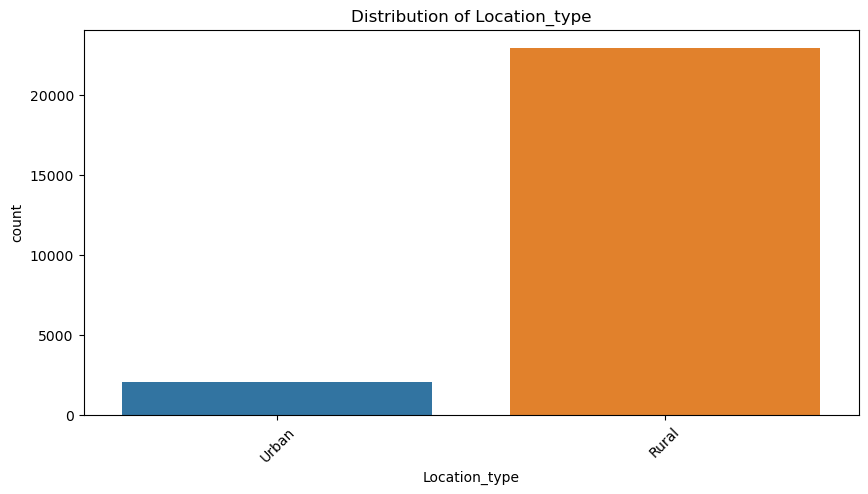
**Exploratory Data Analysis (EDA)** The EDA process was conducted to get a better understanding of the dataset and the relationships between the different variables. EDA is a critical step in the project as it helps uncover valuable insights and relationships within the data. This includes visualizing data, identifying trends, and understanding how different factors affect the supply chain.

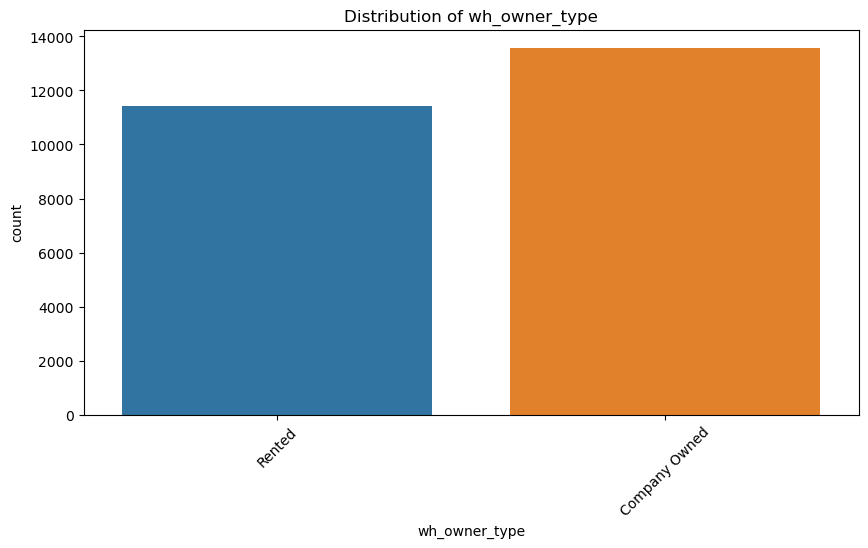
**Data Preprocessing** Data preprocessing involves preparing the data for modeling by encoding categorical variables, normalizing data, and making it ready for regression analysis.

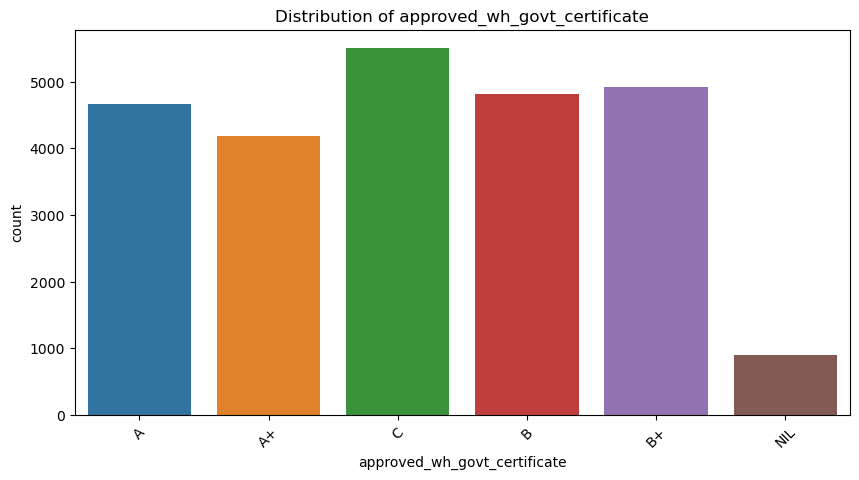
Distribution of the data-



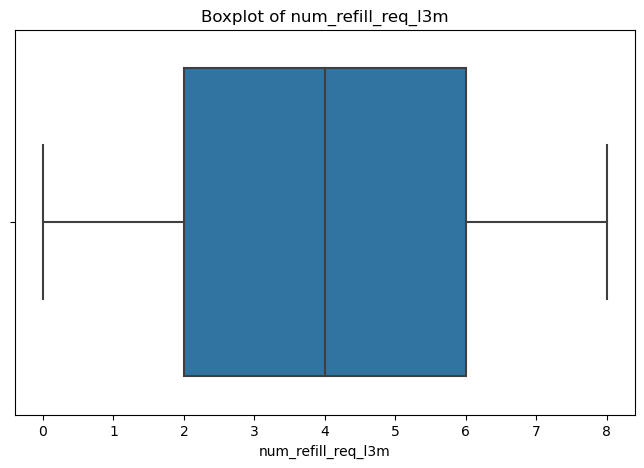
**Univariate Analysis –**

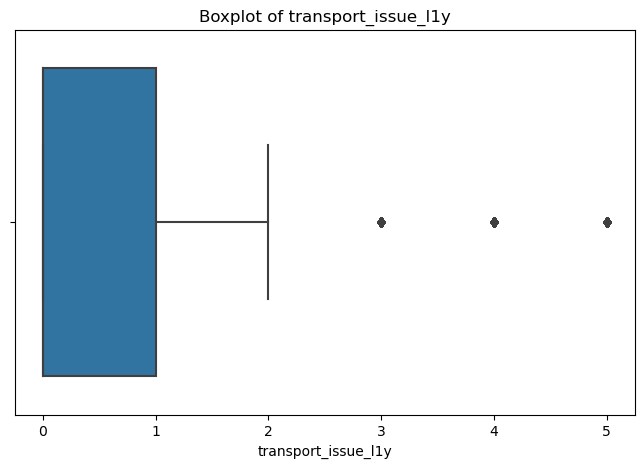


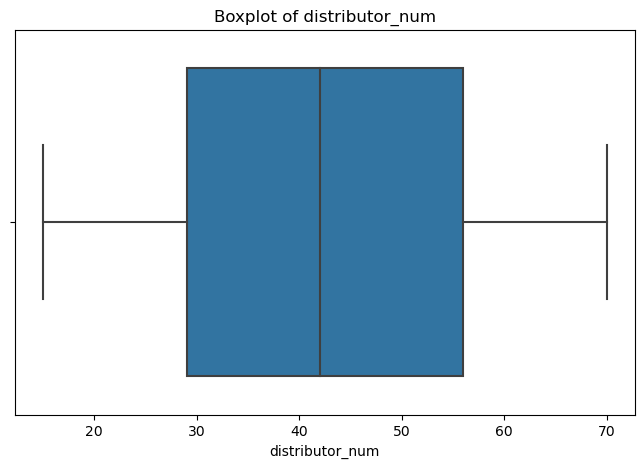
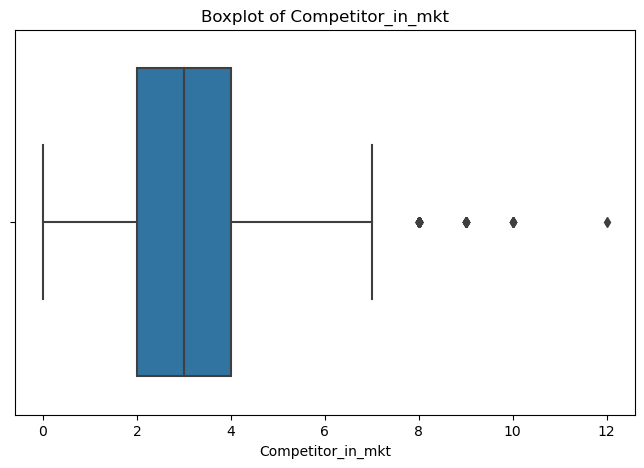


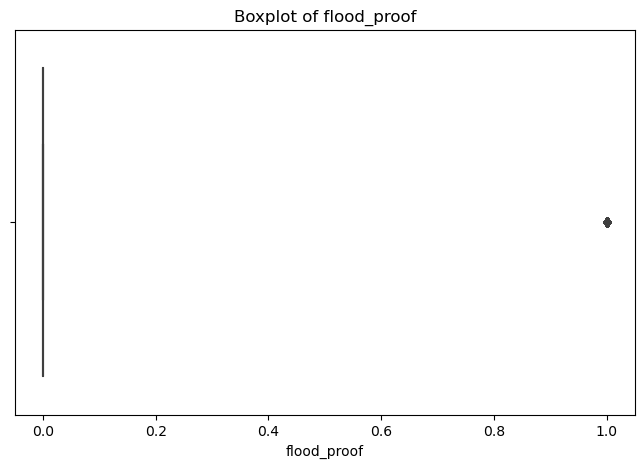


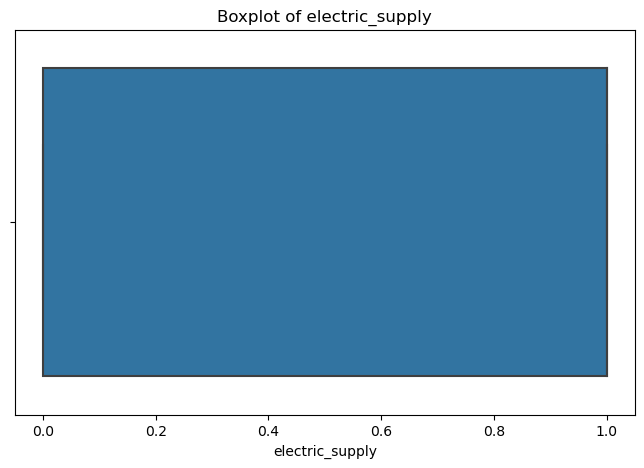
**Univariate Analysis for Continuous features using Boxplots**

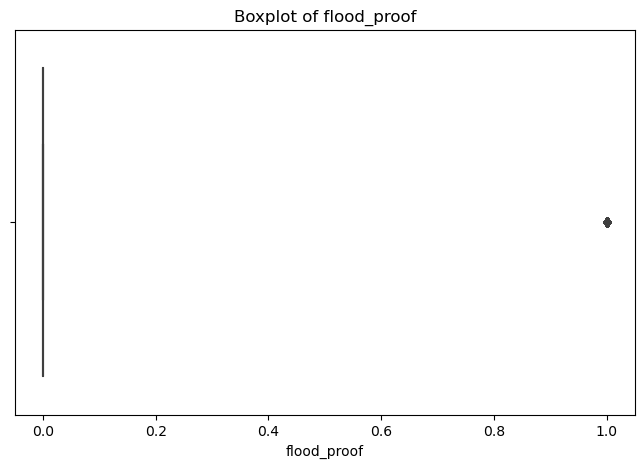


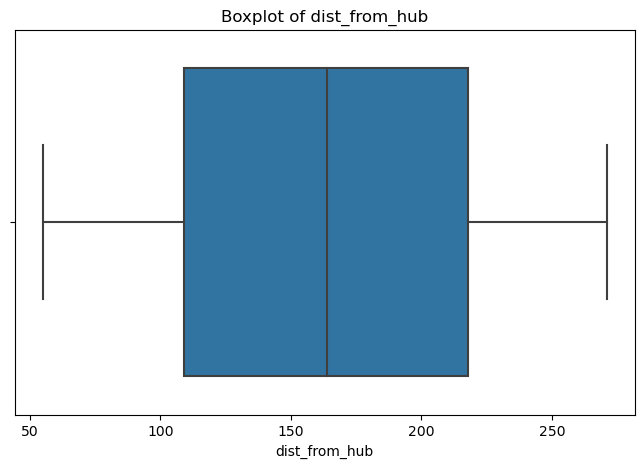
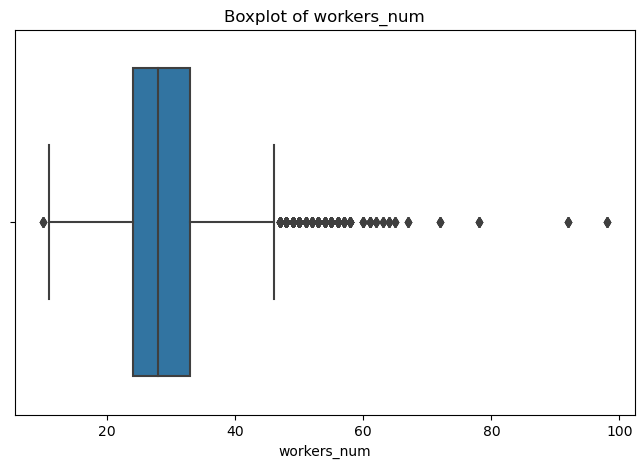


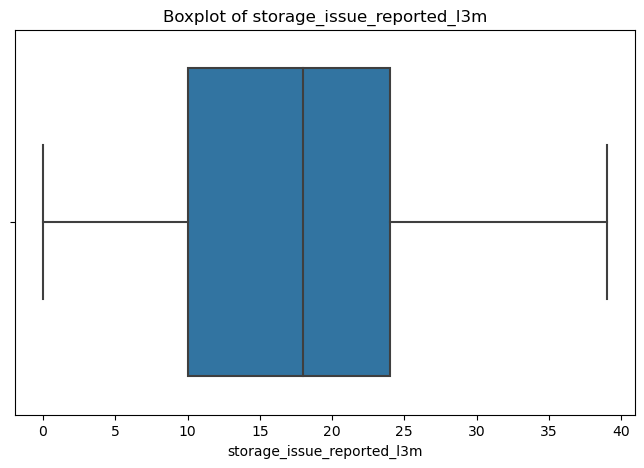
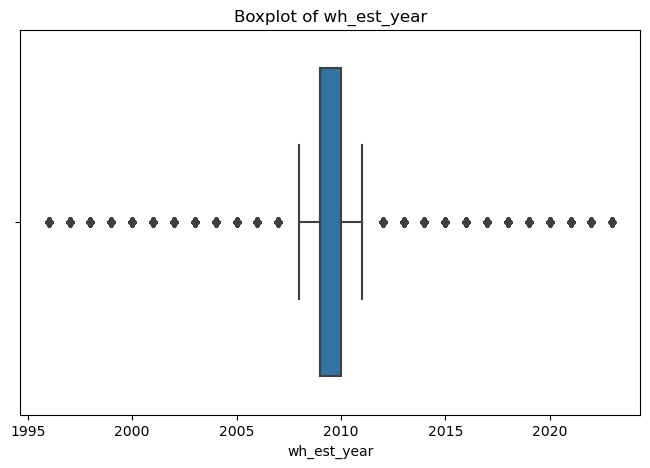


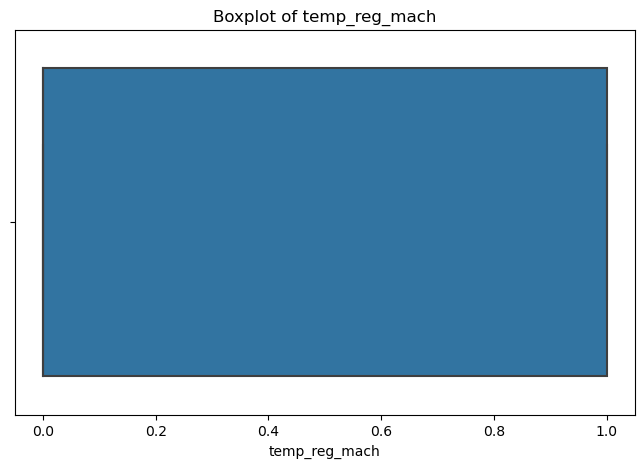


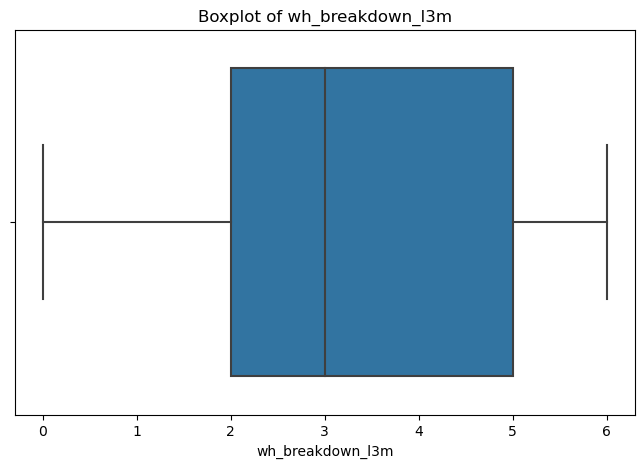


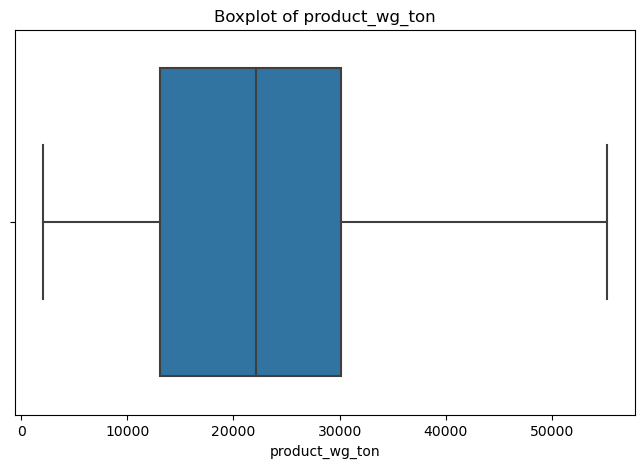




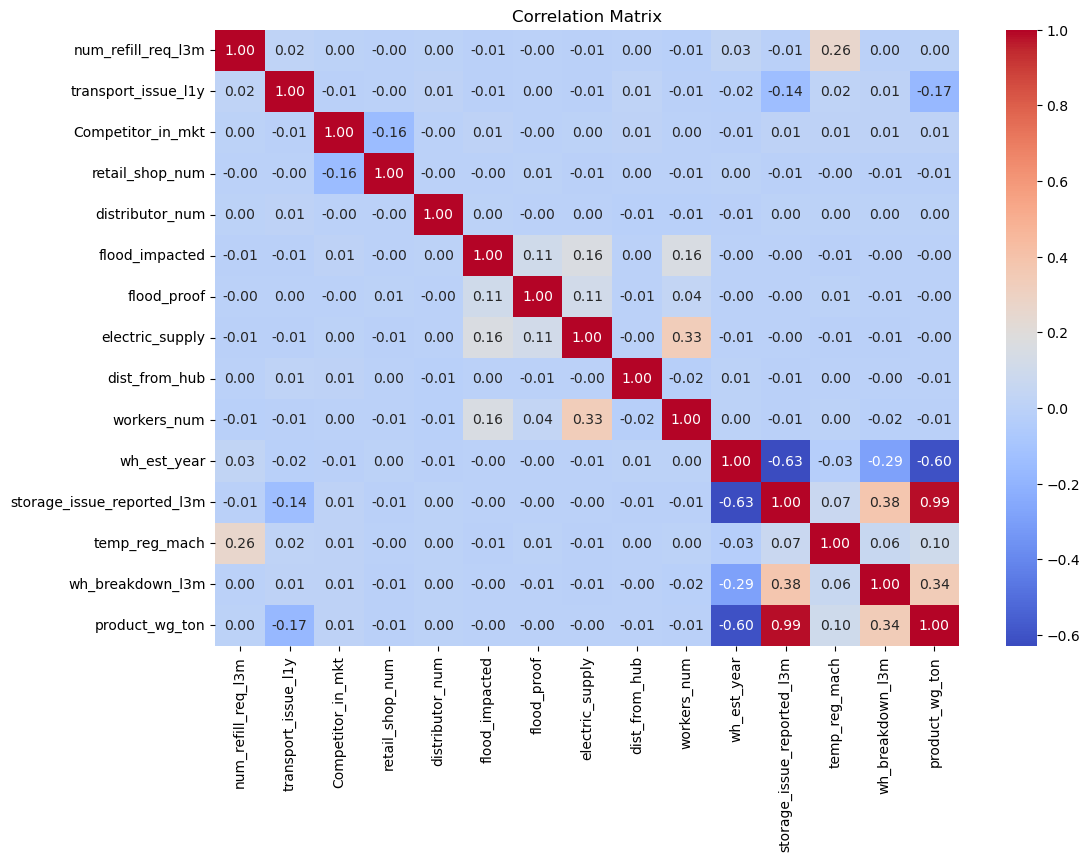




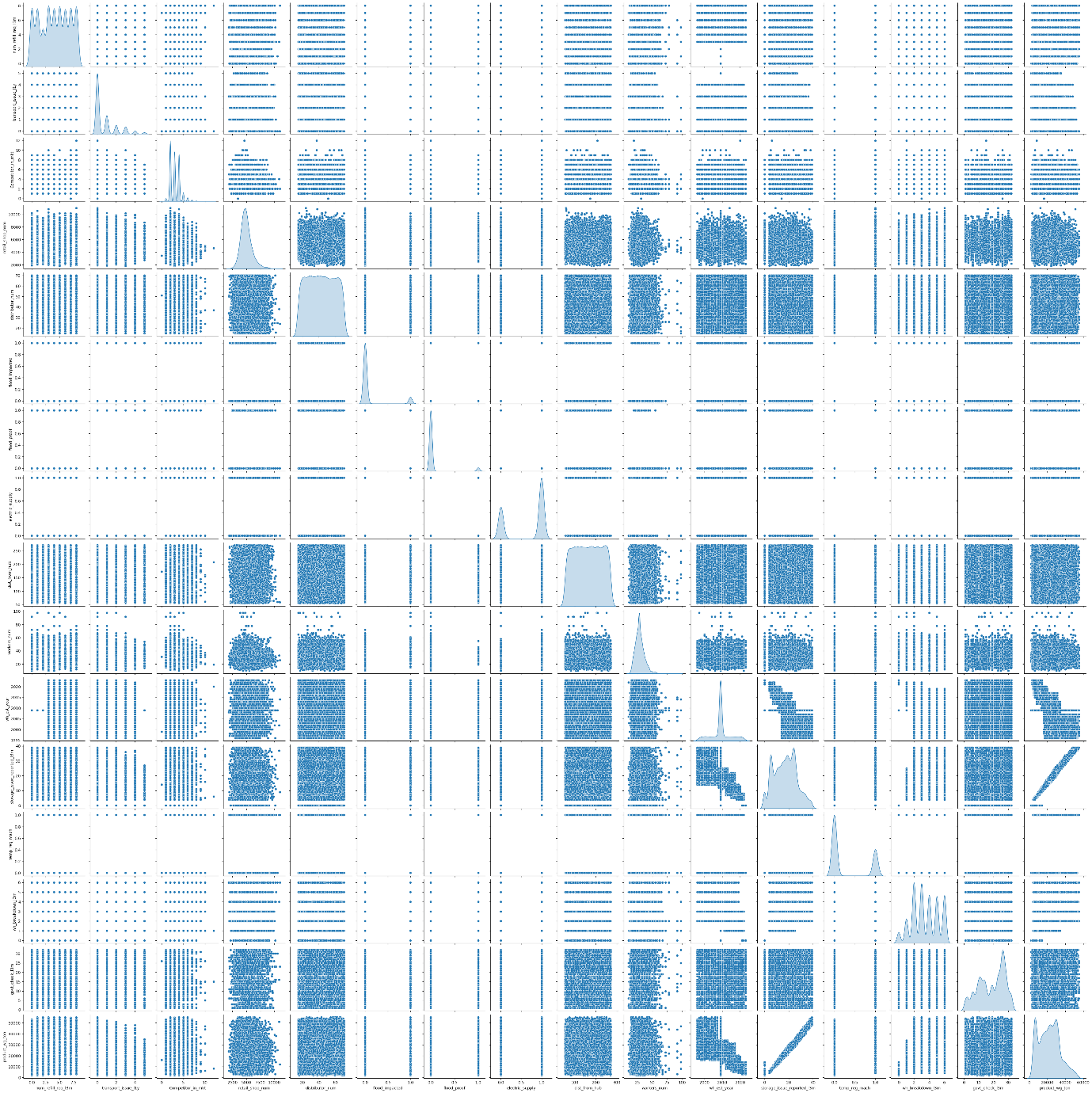




**Bivariate Analysis-**

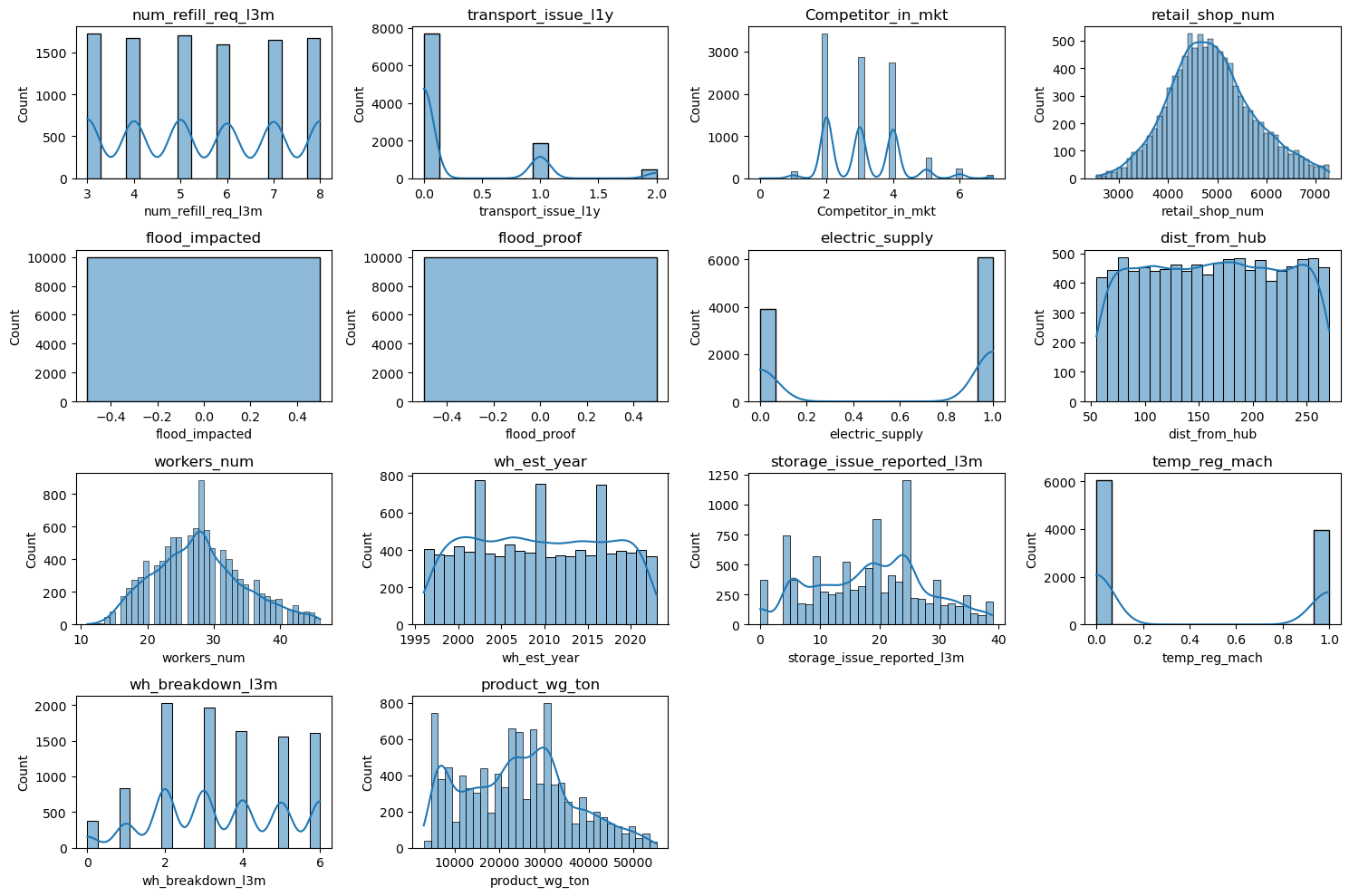


**Multivariate analysis –**

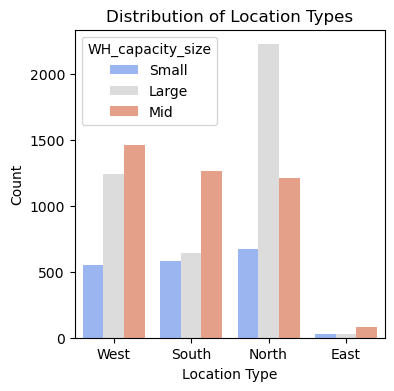
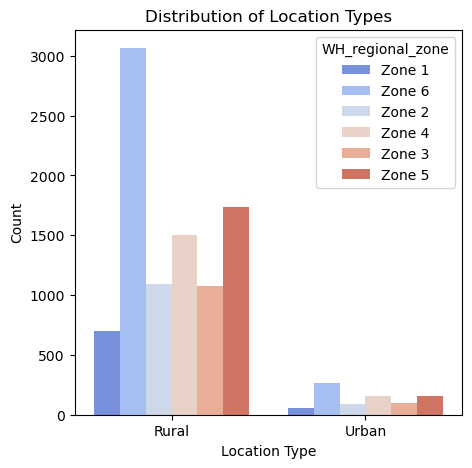


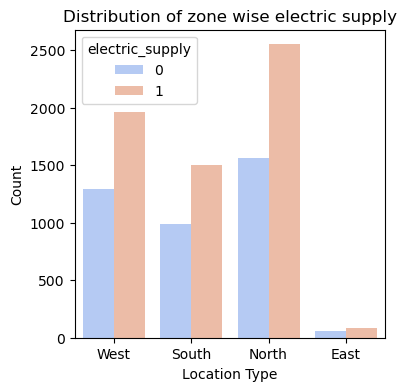
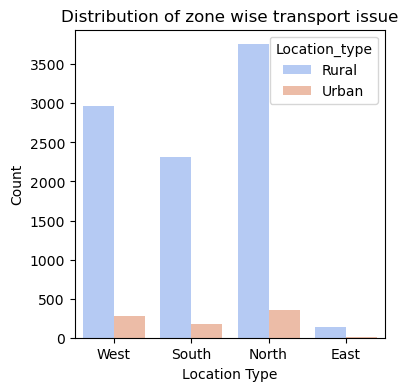
**Data Visualization-**

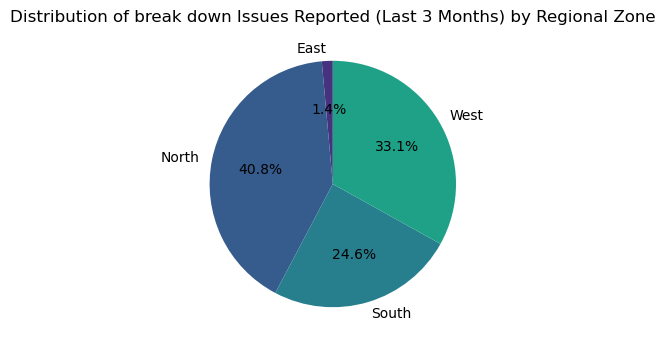
Graphic representation of each variable by histogram

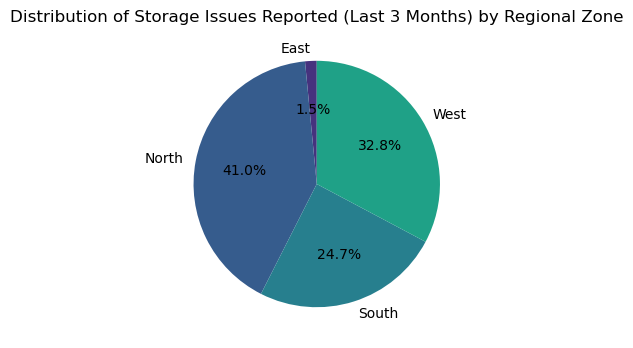
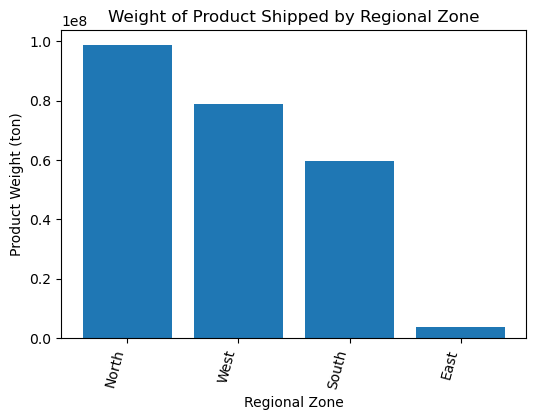


Graphic representation of distribution with loction,regions,transport and electric suplly





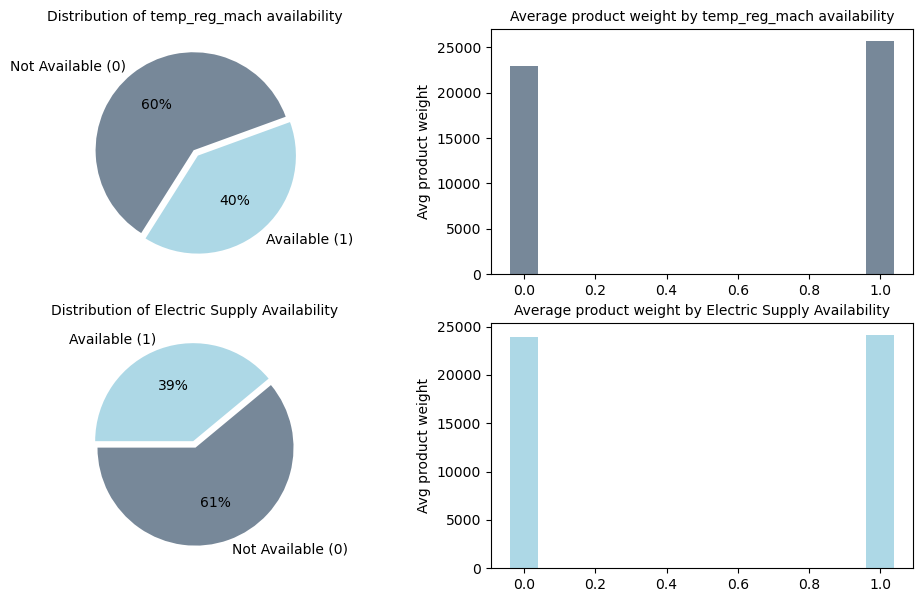
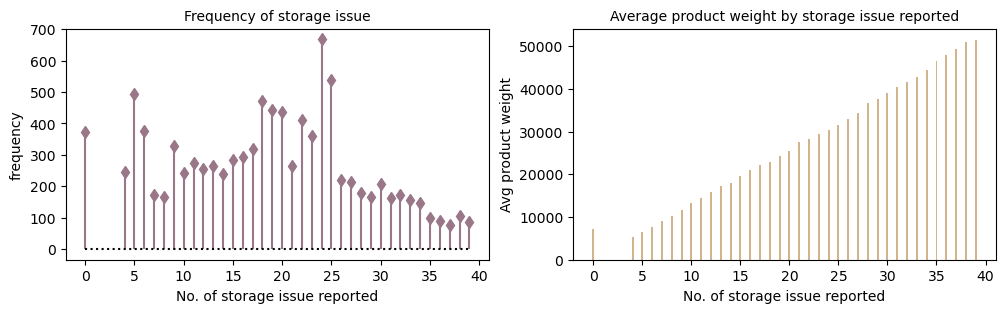
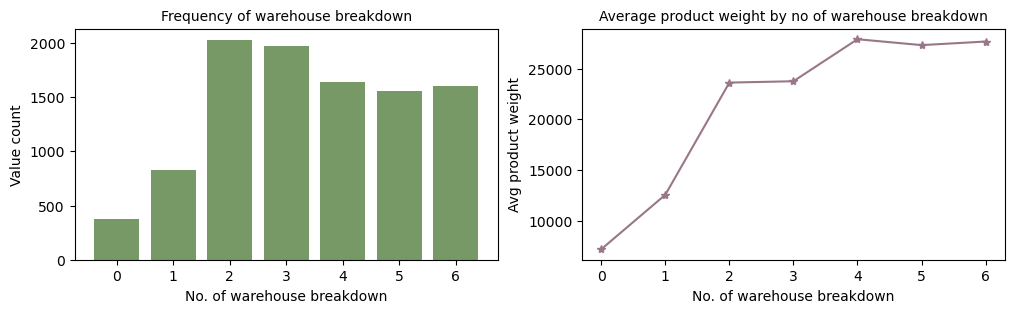


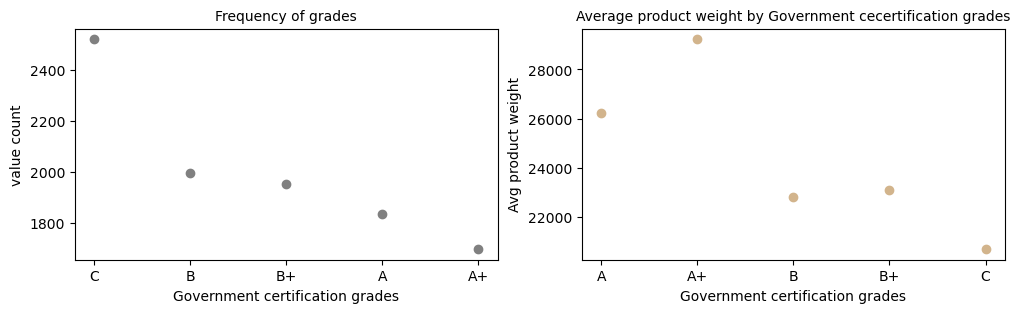
correlation matrix for finding the relationship between two variables

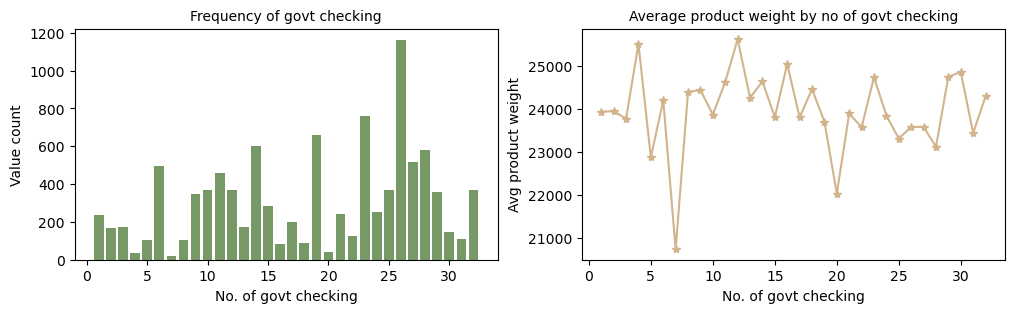
**Warehouse location type & Owner Type**

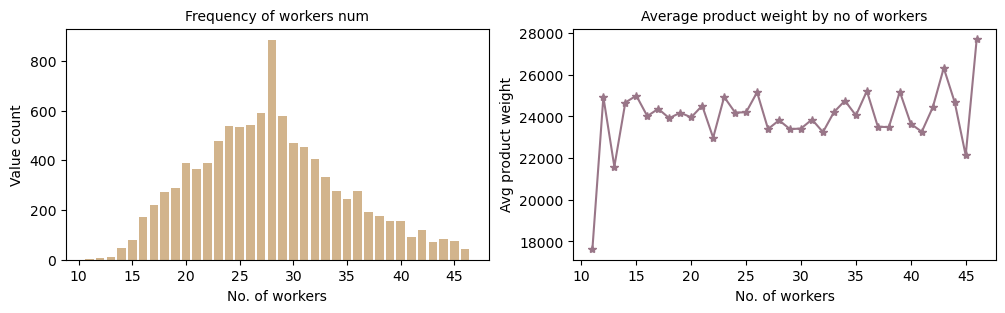
**zone & Regional\_zone**

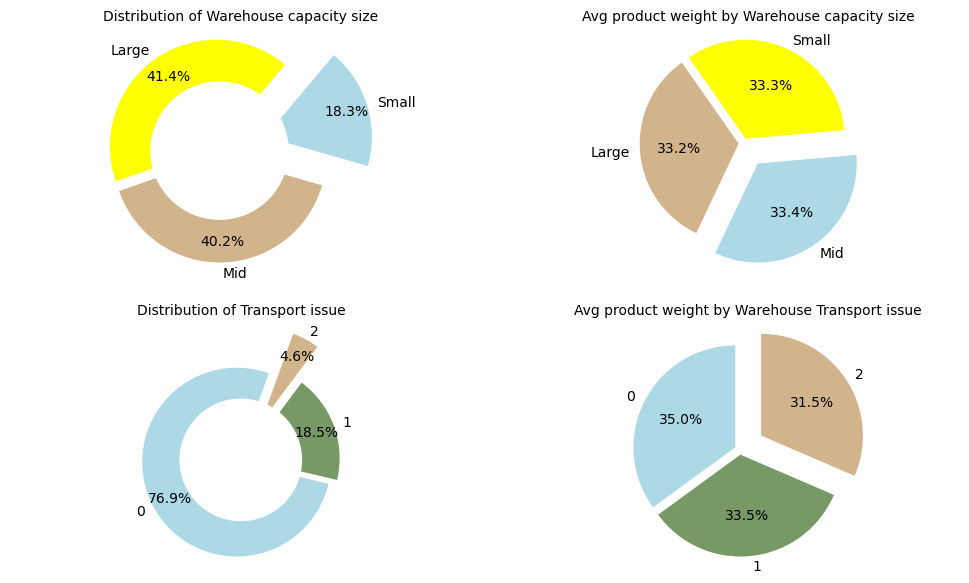
**Electric\_supply & temp\_reg\_mach availability**

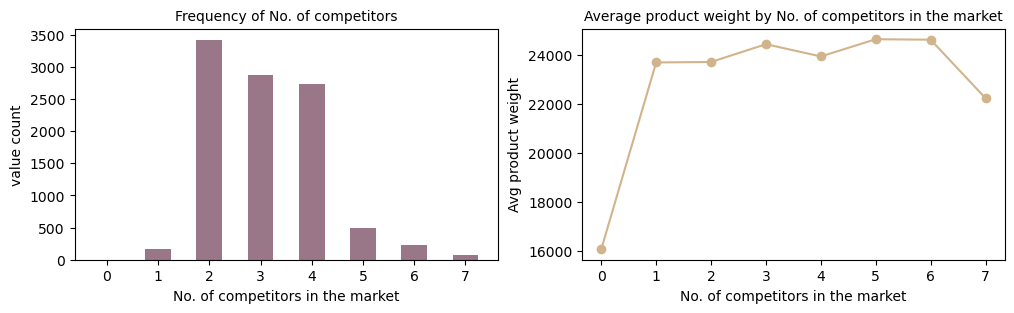
 **Storage issue reported in the last 3 months****No. of warehouse breakdown in the last 3 monthszone & Regional\_zone****Type of approval by government**

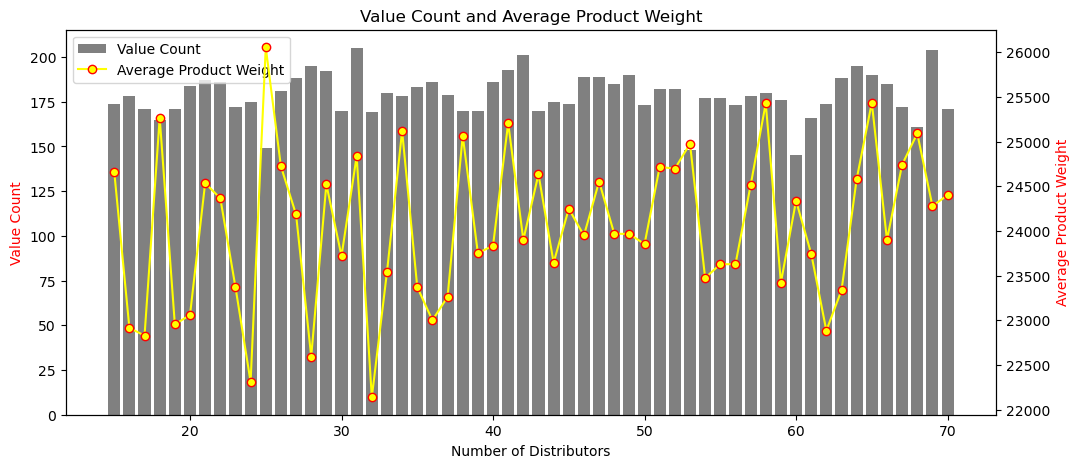
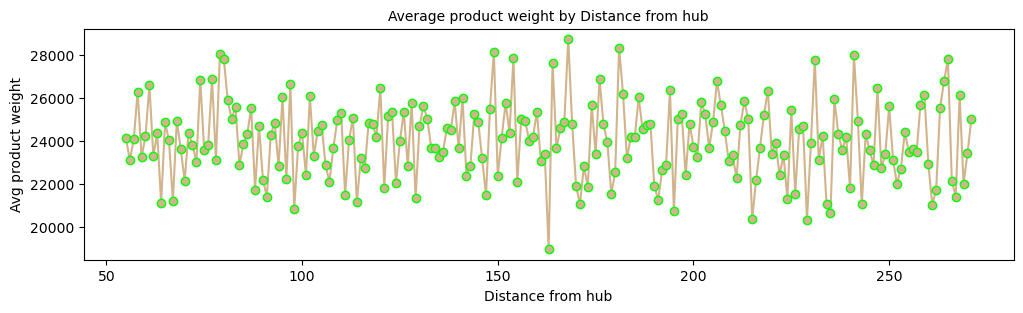
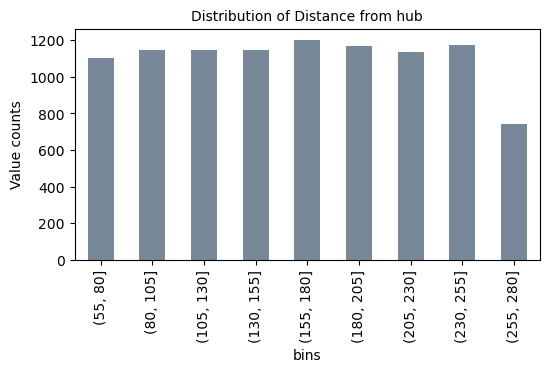
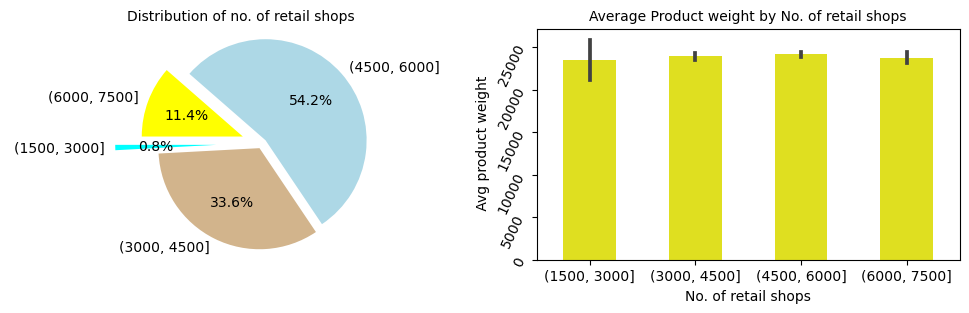
**govt checking in last 3 months**

**No. of workers**

**Warehouse capacity size & Transport issue**

 **Competitor\_in\_mkt**

**No. of Distributors**

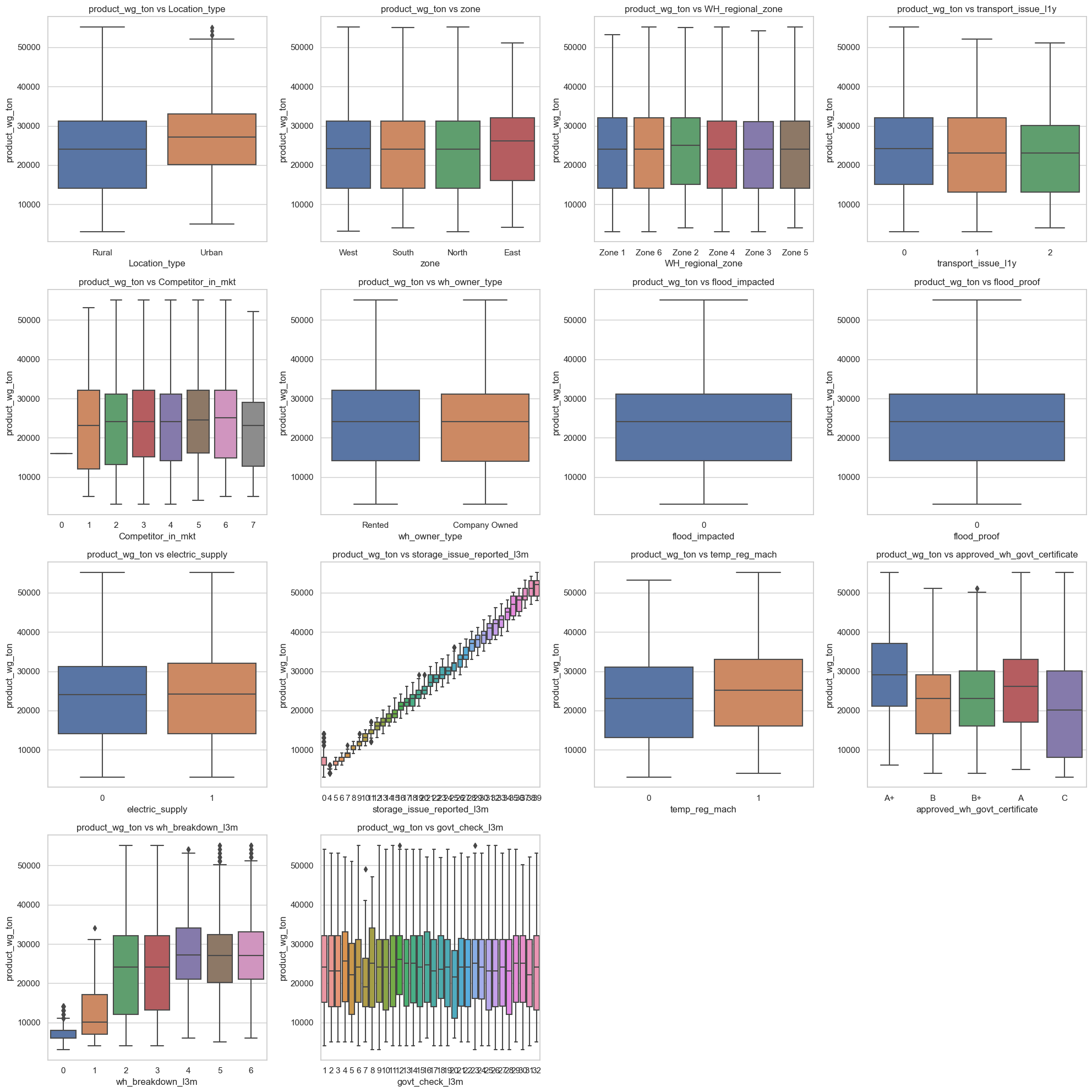
**Distance from hub****Retail shop number**

**Model Building**

The models are evaluated used accuracy score and RMSE values and chose the best model and find out the feature importance that affect the optimum weight for we will have to do again data preprocessing for split the data into test set and train set.

**Data Preprocessing** Preparing the data for modeling by encoding categorical variables, normalizing data, and making it ready for regression analysis.

**Model Building and Evaluation** Here models is built using the following machine learning algorithms & fitted to the training data and evaluated using the testing data.



Created a box plots to compare a numerical variable ('product\_wg\_ton')

**Machine learning model and evaluation**

**Linear /Lasso/Ridge Regression-**

Linear Regression:

Mean Squared Error: 2098749.863484956

Root Mean Squared Error: 1448.7062723288514

Mean Absolute Error: 1121.8516297020822

R-squared Score: 0.9809045125374797

Lasso Regression:

Mean Squared Error: 2099274.4461058103

Root Mean Squared Error: 1448.8873131150713

Mean Absolute Error: 1121.9150828979914

R-squared Score: 0.9808997396195471

Ridge Regression:

Mean Squared Error: 2098757.843358652

Root Mean Squared Error: 1448.7090264641317

Mean Absolute Error: 1121.8525693718934

R-squared Score: 0.9809044399325539

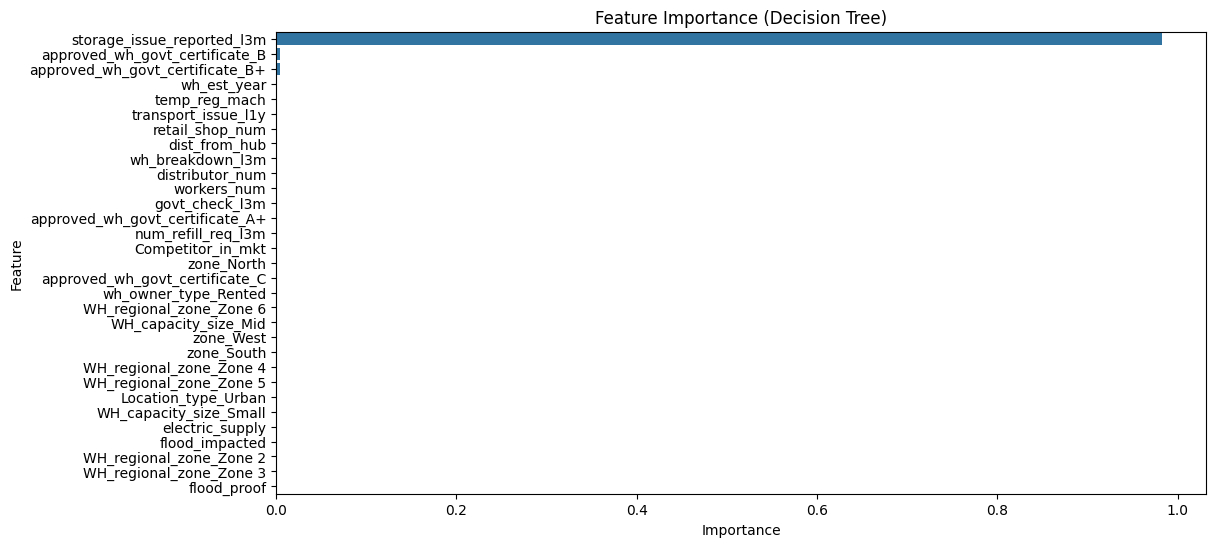
**Decision Tree Regressor-**

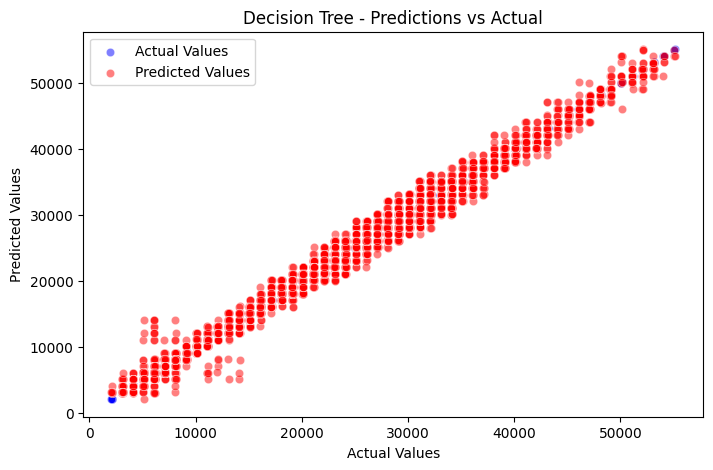
Mean Squared Error: 0.0

Root Mean Squared Error: 0.0

Mean Absolute Error: 0.0

R-squared Score: 1.0





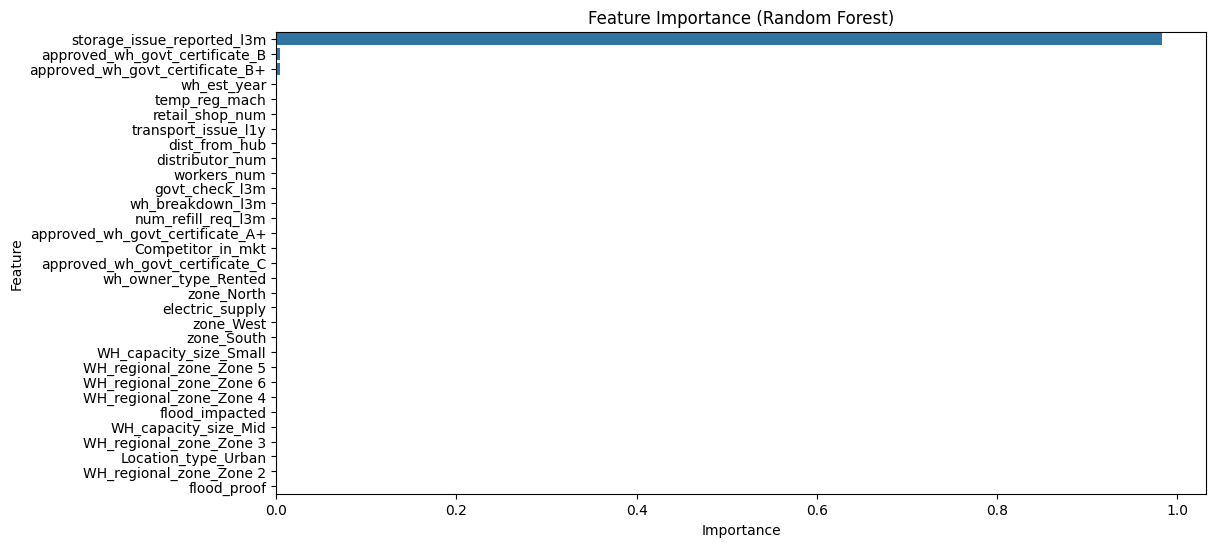
**Random Forest Regressor-**

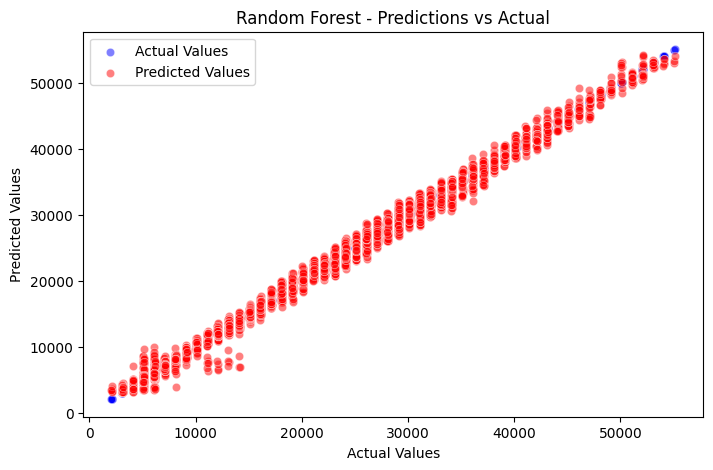
Mean Squared Error: 108548.23710655098

Root Mean Squared Error: 329.4665948264725

Mean Absolute Error: 252.02583080260305

R-squared Score: 0.999012373252854





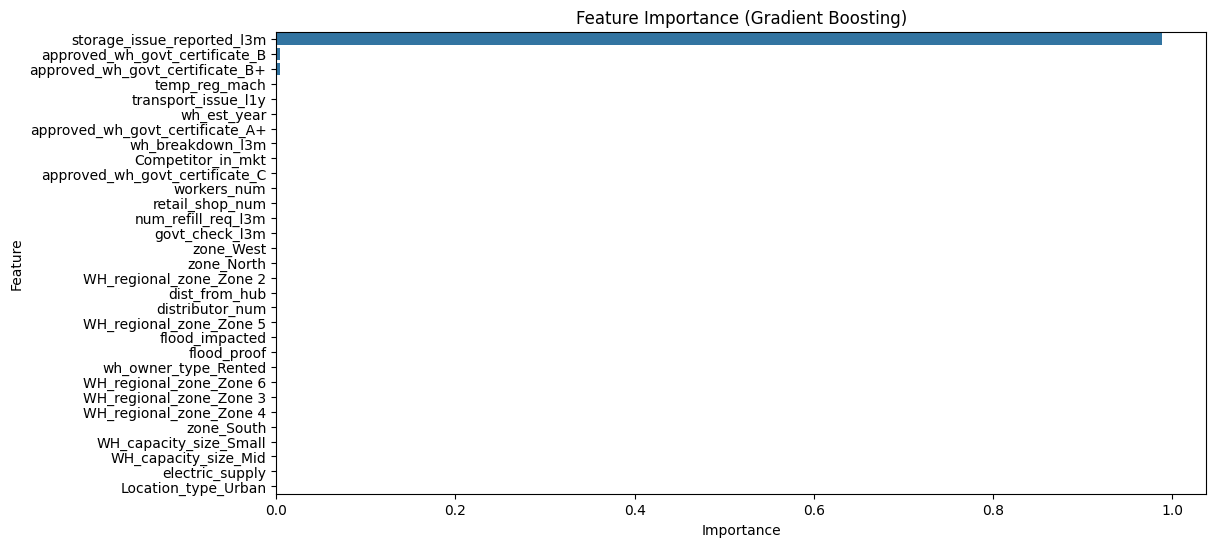
**GradientBoostingRegressor-**

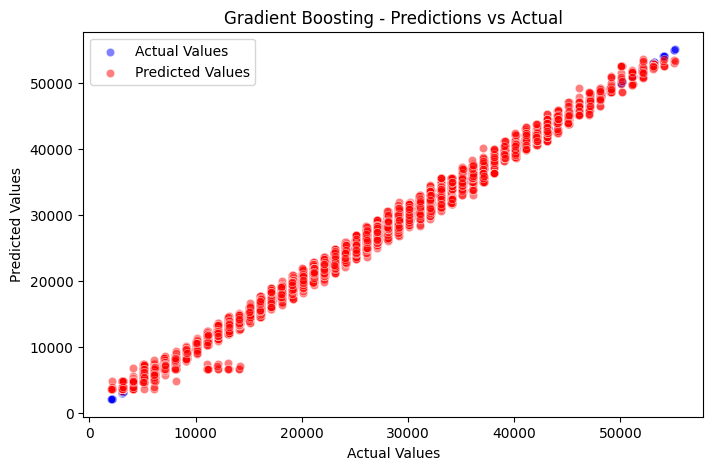
Mean Squared Error: 631885.5437349003

Root Mean Squared Error: 794.9122868184264

Mean Absolute Error: 625.3112301258157

R-squared Score: 0.9942507858187055





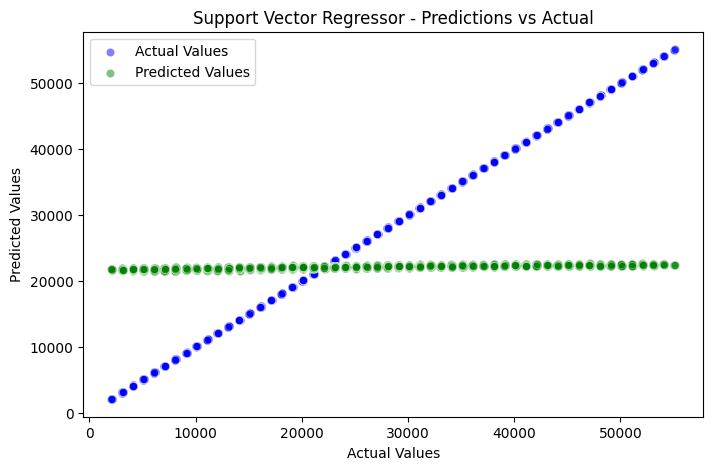
**Support Vector Regression-**

Mean Squared Error: 107076054.56935987

Root Mean Squared Error: 10347.756016130254

Mean Absolute Error: 8307.763091717337

R-squared Score: 0.025767913966559197



**KNeighbors Regressor-**

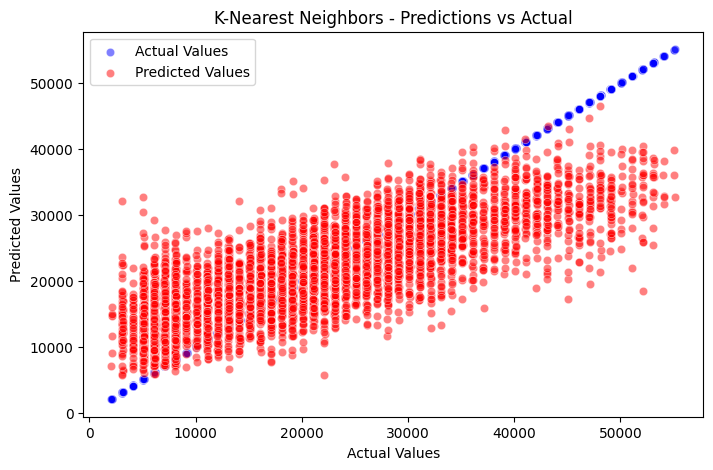
K-Nearest Neighbors:

Mean Squared Error: 40424423.996668115

Root Mean Squared Error: 6358.0204463864475

Mean Absolute Error: 5052.002082429502

R-squared Score: 0.6321981504140726



**Evaluation** – Here regression models is assessed using appropriate evaluation metrics helps determine which model is the most effective in optimizing the supply chain.

**Conclusion-Here supply chain management is supervised by regression analysis & with the** power of data preprocessing, feature engineering, and regression modeling, we can uncover valuable insights and improve the efficiency of the supply chain. The project's findings and models will have practical applications in real-world supply chain optimization.

## Suggestions

## Firstly in the supply chain when storage issue increases. Hence, storage capacity must be upgraded to accomodate higher product weights.It is important process of regularly assessing and expanding storage infrastructure in line with the production increments to ensure smooth warehouse operations and prevent bottlenecks.

## Explore & opportunities to enhance product quality standards across warehouse facilities.

## Examine various factors such as technology adoption, training programs, and workflow optimization strategies to understand their collective impact on production efficiency.

## Minimizing the risk of operational disruptions caused by breakdowns. Conduct comprehensive analysis to identify root causes of breakdowns during periods of heightened production.

## Expands the numbers of warehouse operations in rural areas to capitalize on the higher average product weights and the share of total product weight, which can potentially lead to increase overall productivity and profitability.