CAPSTONE PROJECT

PGP- Data Science and Business Analytics

Abstract

This report analyzes supply chain inefficiencies in a FMCG company's instant noodle business, where demand supply mismatch leads to inventory loss.

Akhil Mohandas

akhilmdas96@gmail.com

Introduction Of The Business Problem

Problem Statement

The FMCG company is facing an issue with the mismatch between demand and supply for their instant noodles across various warehouses in the country. This imbalance is resulting in both excess inventory in some areas and stockouts in others, leading to significant inventory cost losses. The company seeks to optimize the supply chain by determining the ideal quantity of product to ship to each warehouse based on historical demand data. Furthermore, the company wants to analyze regional demand patterns to refine its advertising campaigns and target specific pockets of the country where demand is high but supply is low.

Data Dictionary

File: Data.csv

Target variable: product_wg_ton

Ware_house_ID	Product warehouse ID
WH_Manager_ID	Employee ID of warehouse manager
Location_type	Location of warehouse like in city or village
WH_capacity_size	Storage capacity size of the warehouse
zone	Zone of the warehouse
WH_regional_zone	Regional zone of the warehouse under each zone
num_refill_req_l3m	Number of times refilling has been done in last 3 months
transport_issue_l1y	Any transport issue like accident or goods stolen reported in last one year
Competitor_in_mkt	Number of instant noodles competitor in the market
retail_shop_num	Number of retails shop who sell the product under the warehouse area
wh_owner_type	Company is owning the warehouse or they have get the warehouse on rent
distributor_num	Number of distributer works in between warehouse and retail shops
flood_impacted	Warehouse is in the Flood impacted area indicator
flood_proof	Warehouse is flood proof indicators. Like storage is at some height not directly on the ground
electric_supply	Warehouse have electric backup like generator, so they can run the warehouse in load shedding
dist_from_hub	Distance between warehouse to the production hub in Kms
workers_num	Number of workers working in the warehouse
wh_est_year	Warehouse established year

storage_issue_reported_l3m	Warehouse reported storage issue to corporate office in last 3 months. Like rat, fungus because of moisture etc.
temp_reg_mach	Warehouse have temperature regulating machine indicator
approved_wh_govt_certificate	What kind of standard certificate has been issued to the warehouse from government regulatory body
wh_breakdown_l3m	Number of time warehouse face a breakdown in last 3 months. Like strike from worker, flood, or electrical failure
govt_check_l3m	Number of time government Officers have been visited the warehouse to check the quality and expire of stored food in last 3 months
product_wg_ton	Product has been shipped in last 3 months. Weight is in tons

Need Of The Study/Project

The study aims to develop a data-driven model that:

- **Optimizes the Supply Chain**: The key objective is to accurately predict the optimal quantity of product to be shipped to each warehouse, minimizing both understocking and overstocking scenarios. By doing so, the company will reduce its inventory costs and enhance overall operational efficiency.
- **Improves Demand Forecasting**: The study will also involve analyzing demand patterns across different regions of the country. This will allow the company to understand which areas experience higher demand for instant noodles and at what times, helping to guide better supply decisions.
- **Refines Marketing Strategy**: Based on the demand analysis, the company can optimize its advertisement strategy, ensuring it focuses efforts on regions with the highest potential for growth. Targeting specific pockets of high demand can result in higher sales and improved brand recognition in those areas.
- **Provides Actionable Insights for Phase 2**: The initial success of the model using limited data will serve as a proof of concept, paving the way for access to a more comprehensive data lake in the future, which can be used to develop a more sophisticated and robust supply chain optimization model.

Business/Social Opportunity

1.Business Opportunity:

a) **Cost Reduction**: Efficient supply chain management can lead to significant cost savings. By minimizing excess inventory and avoiding stockouts, the company can optimize storage costs, transportation costs, and reduce wastage.

b) **Revenue Maximization**: A more accurate supply system ensures products are available where and when they are needed, thus increasing sales opportunities and customer satisfaction. By analyzing regional demand, the company can also tailor marketing efforts to boost sales in high-demand areas.

c) **Competitive Advantage**: The company can gain an edge over competitors by efficiently managing its supply chain, ensuring they are more responsive to market conditions and consumer demands.

2.Social Opportunity:

a) **Increased Product Accessibility**: By optimizing supply, the company ensures that their products are available to consumers where they are needed most, improving accessibility in both urban and rural areas.

b) **Economic Growth**: Efficient supply chains contribute to the overall economic development of local areas by enabling more consistent availability of goods. This can indirectly support regional economic stability and employment.

c) **Sustainability**: By minimizing wastage through optimized supply, the company can take a more sustainable approach to inventory management, which can improve its brand image and align with global environmental goals.

Data Report

Visual Inspection Of Data

After importing the file containing the data, we read the first few columns(first 5 columns).

)	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	...
0	EID_50000	Urban	Small	West	Zone 6	3	1	2	4651	...
1	EID_50001	Rural	Large	North	Zone 5	0	0	4	6217	...
2	EID_50002	Rural	Mid	South	Zone 2	1	0	4	4306	...
3	EID_50003	Rural	Mid	North	Zone 3	7	4	2	6000	...
4	EID_50004	Rural	Large	North	Zone 5	3	1	2	4740	...

Table 1: Top 5 rows of Dataset

Shape of the data:

The data contains 25000 number of rows or observations and 24 columns or variables.

Basic Info of the data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Ware_house_ID                  25000 non-null  object
 1   WH_Manager_ID                  25000 non-null  object
 2   Location_type                  25000 non-null  object
 3   WH_capacity_size               25000 non-null  object
 4   zone                           25000 non-null  object
 5   WH_regional_zone               25000 non-null  object
 6   num_refill_req_l3m             25000 non-null  int64
 7   transport_issue_l1y            25000 non-null  int64
 8   Competitor_in_mkt              25000 non-null  int64
 9   retail_shop_num                25000 non-null  int64
 10  wh_owner_type                  25000 non-null  object
 11  distributor_num                25000 non-null  int64
 12  flood_impacted                 25000 non-null  int64
 13  flood_proof                    25000 non-null  int64
 14  electric_supply                25000 non-null  int64
 15  dist_from_hub                  25000 non-null  int64
 16  workers_num                    24010 non-null  float64
 17  wh_est_year                    13119 non-null  float64
 18  storage_issue_reported_l3m     25000 non-null  int64
 19  temp_reg_mach                  25000 non-null  int64
 20  approved_wh_govt_certificate   24092 non-null  object
 21  wh_breakdown_l3m               25000 non-null  int64
 22  govt_check_l3m                 25000 non-null  int64
 23  product_wg_ton                 25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Table 2: Basic Info of the dataset

The dataset consists of 25,000 warehouse records with 24 attributes.Some columns, such as *approved_wh_govt_certificate*,*workers_num* and *wh_est_year*, have missing values (*workers_num*: 990 missing, *wh_est_year*: 11,881 missing,*approved_wh_govt_certificate:*908), which may impact analysis.Data types are mostly appropriate, with a mix of categorical (*object*), numerical (*int64*, *float64*), and binary indicators.

16 Numerical columns and 8 categorical columns.The "*wh_est_year*" column has significant missing data, which might affect analysis on older vs. newer warehouses. "*workers_num*" and "*approved_wh_govt_certificate*" have relatively fewer missing values and can be imputed.

Statistical Summary:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ware_house_ID	25000	25000	WH_100000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_Manager_ID	25000	25000	EID_50000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Location_type	25000	2	Rural	22957	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_capacity_size	25000	3	Large	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zone	25000	4	North	10278	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_regional_zone	25000	6	Zone 6	8339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num_refill_req_l3m	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	NaN	NaN	NaN	4985.71156	1052.825252	1821.0	4313.0	4859.0	5500.0	11008.0
wh_owner_type	25000	2	Company Owned	13578	NaN	NaN	NaN	NaN	NaN	NaN	NaN
distributor_num	25000.0	NaN	NaN	NaN	42.41812	16.064329	15.0	29.0	42.0	56.0	70.0
flood_impacted	25000.0	NaN	NaN	NaN	0.09816	0.297537	0.0	0.0	0.0	0.0	1.0
flood_proof	25000.0	NaN	NaN	NaN	0.05464	0.227281	0.0	0.0	0.0	0.0	1.0
electric_supply	25000.0	NaN	NaN	NaN	0.65688	0.474761	0.0	0.0	1.0	1.0	1.0
dist_from_hub	25000.0	NaN	NaN	NaN	163.53732	62.718609	55.0	109.0	164.0	218.0	271.0
workers_num	24010.0	NaN	NaN	NaN	28.944398	7.872534	10.0	24.0	28.0	33.0	98.0
wh_est_year	13119.0	NaN	NaN	NaN	2009.383185	7.52823	1996.0	2003.0	2009.0	2016.0	2023.0
storage_issue_reported_l3m	25000.0	NaN	NaN	NaN	17.13044	9.161108	0.0	10.0	18.0	24.0	39.0
temp_reg_mach	25000.0	NaN	NaN	NaN	0.30328	0.459684	0.0	0.0	0.0	1.0	1.0
approved_wh_govt_certificate	24092	5	C	5501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wh_breakdown_l3m	25000.0	NaN	NaN	NaN	3.48204	1.690335	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	NaN	NaN	NaN	18.81228	8.632382	1.0	11.0	21.0	26.0	32.0
product_wg_ton	25000.0	NaN	NaN	NaN	22102.63292	11607.755077	2065.0	13059.0	22101.0	30103.0	55151.0

Table 3: Statistical Summary of Dataset

Distribution and Trends:

1. Warehouse Locations and Capacities:
    - Most warehouses are in rural areas.
    - Warehouses are mostly Large followed by medium and small.
    - The highest number of warehouses are in the North zone.
2. Retail Shop and Distributor Network:
    - Average number of retail shops served per warehouse: 4,986.
    - Number of distributors per warehouse: 42.
    - On average, 3 competitors per region, but some areas have up to 12 competitors.
    - Some warehouses serve up to 11,008 retail shops, indicating high demand while the lowest value is 1821.

3. Logistics and Infrastructure:

- Average distance between warehouse to the production hub in Kms: 164 km, with some warehouses as far as 271 km.
- Warehouses farther from hubs might face transport delays and inventory shortages.
- Warehouses faced an average of 3.48 breakdowns in 3 months.
- Warehouses were inspected an average of 18.8 times in 3 months, ensuring regulatory compliance.

4. Demand-Supply Mismatch Insights:

- Some warehouses required up to 8 refills in the last 3 months, while others needed none.
- This suggests uneven demand, leading to overstocking in some areas and shortages in others.
- Average: 22,102 tons, with a range from 2,065 to 55,151 tons.
- A large variation in shipped weight suggests inefficiencies in supply allocation.

5. Flood and Electrical Resilience:

- 9.8% of warehouses are in flood-prone areas, but only 5.5% are flood-proof.
- Only 65.7% of warehouses have electric backup, meaning that 34.3% may face operational delays during power cuts.

Renaming Variables:

To avoid confusion when interpreting analysis results or visualizations we are replacing the name of few columns.More intuitive column names make it easier for business stakeholders to understand the data.This renaming step is a critical part of data preparation, ensuring clean and well-structured data for subsequent analysis and reporting.

List of renamed variables:

- Warehouse_ID (formerly *Ware_house_ID*)
- Warehouse_Manager_ID (formerly *WH_Manager_ID*)
- Warehouse_Ownership (formerly *wh_owner_type*)
- Product_Shipped_Tons (formerly *product_wg_ton*)

```
Index(['Warehouse_ID', 'Warehouse_Manager_ID', 'Location_type',
       'WH_capacity_size', 'zone', 'WH_regional_zone', 'num_refill_req_l3m',
       'transport_issue_l1y', 'Competitor_in_mkt', 'retail_shop_num',
       'Warehouse_Ownership', 'distributor_num', 'flood_impacted',
       'flood_proof', 'electric_supply', 'dist_from_hub', 'workers_num',
       'wh_est_year', 'storage_issue_reported_l3m', 'temp_reg_mach',
       'approved_wh_govt_certificate', 'wh_breakdown_l3m', 'govt_check_l3m',
       'Product_Shipped_Tons'],
      dtype='object')
```

Table 4: Column Names After Renaming

Data Report

Univariate Analysis:

Univariate analysis examines a single variable to identify patterns and summarize key statistics like mean, median, mode, and standard deviation. It helps in understanding data distribution and variability. For example, counting the number of boys and girls in a classroom is a univariate analysis. This method is essential for identifying trends before conducting deeper analysis.Below the distribution of Numerical values are given:

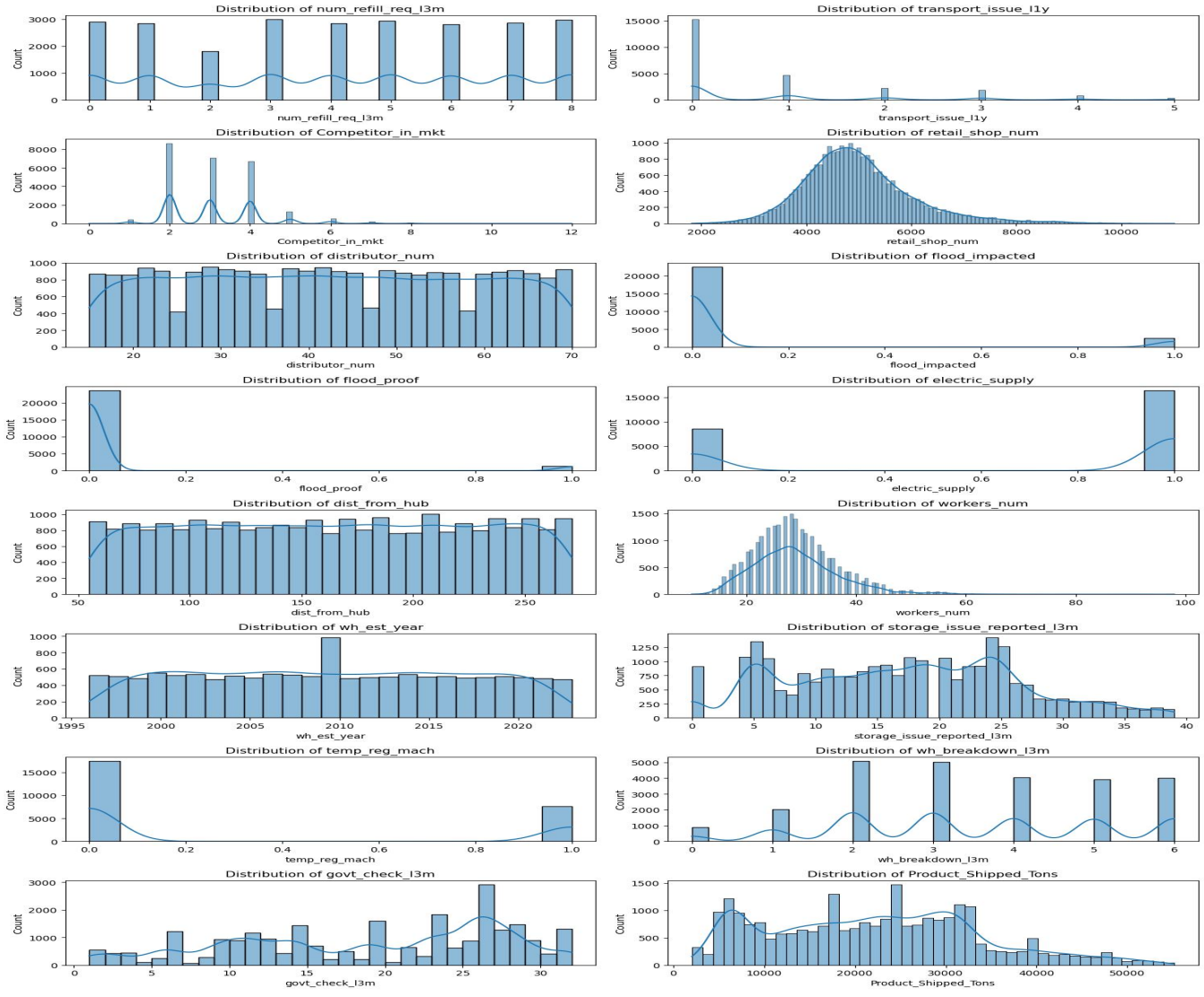

Figure 1: Distribution of Numerical Values

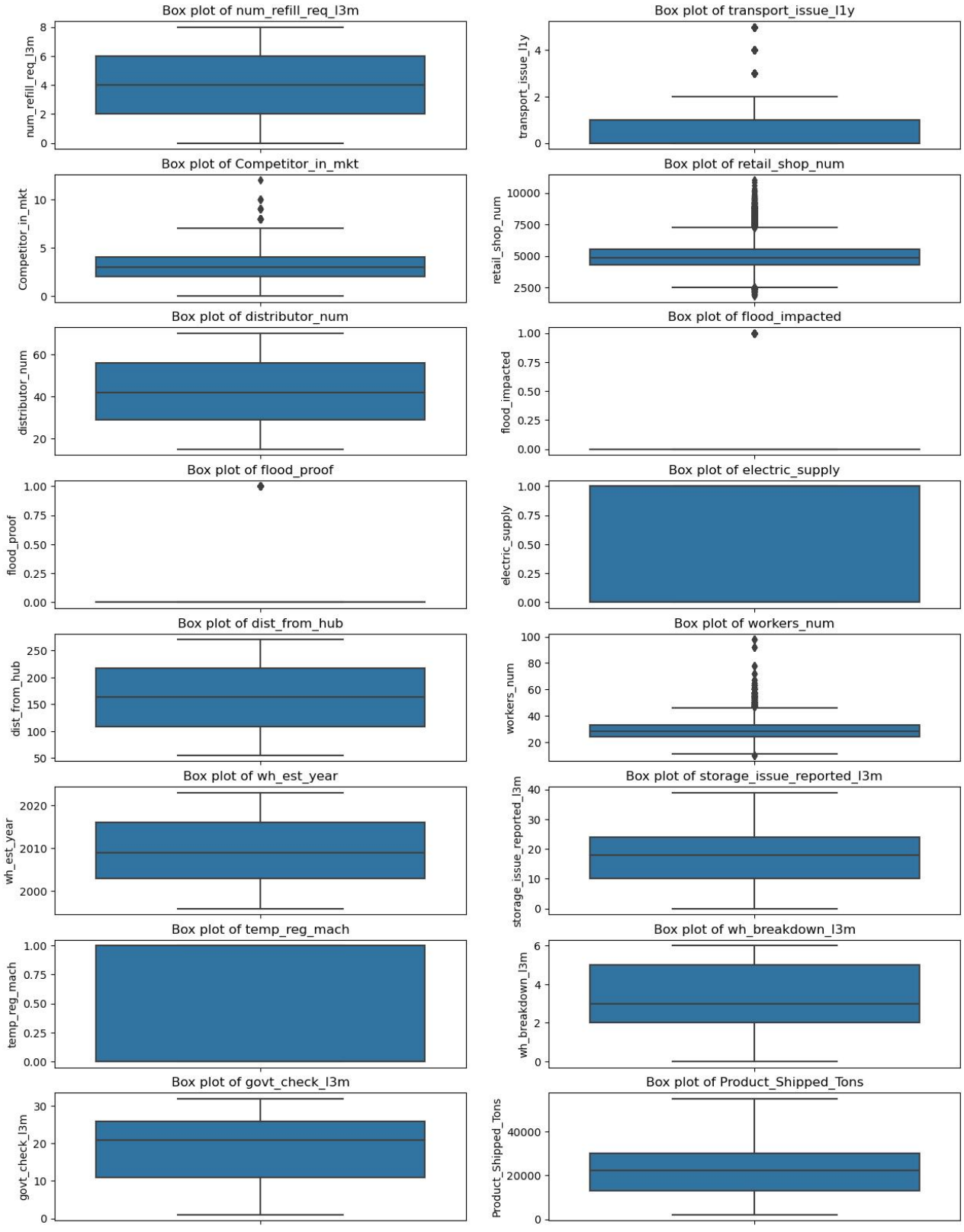

Figure 2: Box Plot of Numerical Values

The table provided below contains statistical insights into key numerical variables related to warehouse operations. It includes metrics like count, mean, standard deviation, minimum, and maximum values.

	count	mean	std	min	25%	50%	75%	max
num_refill_req_l3m	25000.0	4.089040	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	0.773680	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	3.104200	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	4985.711560	1052.825252	1821.0	4313.0	4859.0	5500.0	11008.0
distributor_num	25000.0	42.418120	16.064329	15.0	29.0	42.0	56.0	70.0
flood_impacted	25000.0	0.098160	0.297537	0.0	0.0	0.0	0.0	1.0
flood_proof	25000.0	0.054640	0.227281	0.0	0.0	0.0	0.0	1.0
electric_supply	25000.0	0.656880	0.474761	0.0	0.0	1.0	1.0	1.0
dist_from_hub	25000.0	163.537320	62.718609	55.0	109.0	164.0	218.0	271.0
workers_num	24010.0	28.944398	7.872534	10.0	24.0	28.0	33.0	98.0
wh_est_year	13119.0	2009.383185	7.528230	1996.0	2003.0	2009.0	2016.0	2023.0
storage_issue_reported_l3m	25000.0	17.130440	9.161108	0.0	10.0	18.0	24.0	39.0
temp_reg_mach	25000.0	0.303280	0.459684	0.0	0.0	0.0	1.0	1.0
wh_breakdown_l3m	25000.0	3.482040	1.690335	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	18.812280	8.632382	1.0	11.0	21.0	26.0	32.0
Product_Shipped_Tons	25000.0	22102.632920	11607.755077	2065.0	13059.0	22101.0	30103.0	55151.0

Table 5: Summary of Key Numerical Variables

Observations:

1. **Refill Requests**: Warehouses request an average of 4 refills in the last 3 months, with a maximum of 8. This indicates variability in inventory turnover.**Recommendation**: Optimize stock levels and supply chain management to reduce frequent refill requirements and ensure operational efficiency.
2. **Transport Issues**: Most warehouses report 0 or 1 transport issue per year, but some experience frequent problems. **Recommendation**: Identify and address common causes of transport delays, such as poor road conditions, supplier inefficiencies, or vehicle availability.
3. **Competitors in Market**: Each warehouse operates in a market with an average of 3 competitors, with a maximum of 12. **Recommendation**: Competitive analysis and strategic positioning can help warehouses differentiate their offerings and gain market share.
4. **Retail Shops & Distributors**: Warehouses serve approximately 4,986 retail shops and work with an average of 42 distributors. **Recommendation**: Strengthening relationships with distributors and retailers can enhance efficiency in the supply chain and expand the customer base.
5. **Flood Impact & Proofing**: Only 5.5% of warehouses are flood-proofed, while 9.8% have experienced flood impact. **Recommendation**: Investing in flood-resistant infrastructure and mitigation strategies (e.g., elevated storage, drainage systems) is crucial to reducing potential operational disruptions.
6. **Electric Supply**: Around 66% of warehouses have electricity, leaving a significant portion without reliable power. **Recommendation**: Improving power infrastructure, such as installing solar panels or backup generators, can enhance operational efficiency and prevent product spoilage.
7. **Distance from Hub**: Warehouses are located 164 km away on average, with some as far as 271 km. **Recommendation**: Reducing the distance to distribution hubs, where feasible, can help minimize transportation costs and improve delivery speed.
8. **Storage Issues**: Warehouses report an average of 17 storage issues in the last 3 months, which could indicate capacity constraints or inefficiencies. **Recommendation**: Regular audits and better warehouse management systems can help reduce storage problems.

9. **Breakdowns & Govt Checks**: On average, warehouses experience 3.5 equipment breakdowns and 18 government inspections in the last 3 months. **Recommendation**: Preventive maintenance programs and compliance training can reduce breakdowns and ensure regulatory adherence.

10. **Product Volume**: Warehouses handle an average of 22,100 tons of product shipments, with some moving up to 55,151 tons. **Recommendation**: Optimizing logistics and storage efficiency can help manage high shipment volumes effectively.

11. **Warehouse Establishment Trends**: Most warehouses were established during 2009-2010, suggesting a potential need for modernization and upgrades. **Recommendation**: Assessing older warehouses for structural integrity and upgrading outdated facilities can enhance efficiency and safety.

12. **Temperature Regulation Machines**: Very few warehouses have temperature regulation systems, which could impact storage for perishable goods. **Recommendation**: Investing in temperature-controlled storage can improve product quality and enable entry into new markets, such as pharmaceuticals or fresh produce.

13. **Outliers in Data**: Certain variables show extreme values, suggesting inconsistencies or operational challenges. **Recommendation**: Further investigation is needed to determine whether these outliers represent true business conditions or data inconsistencies, which could impact decision-making.

Below Univariate analysis categorical values are given:

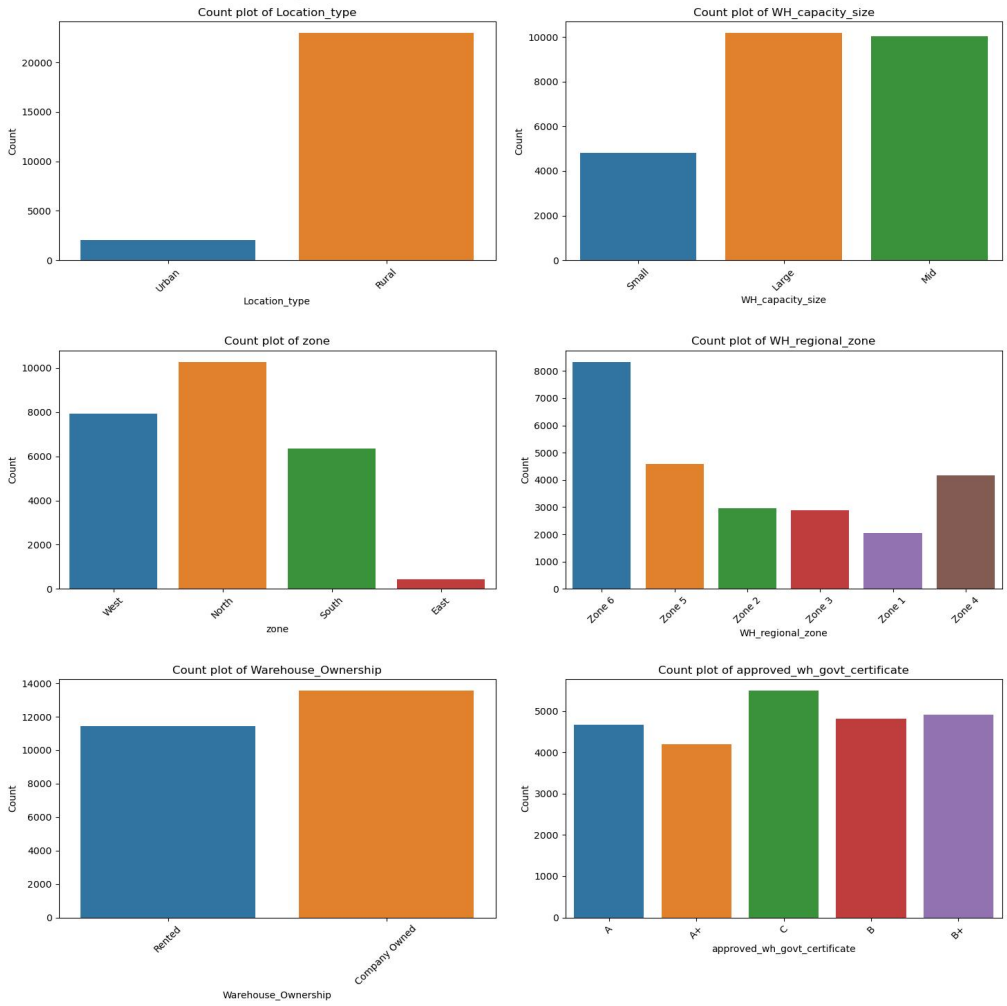

Figure 3: Categorical Variables Countplot

	count	unique	top	freq
Warehouse_ID	25000	25000	WH_100000	1
Warehouse_Manager_ID	25000	25000	EID_50000	1
Location_type	25000	2	Rural	22957
WH_capacity_size	25000	3	Large	10169
zone	25000	4	North	10278
WH_regional_zone	25000	6	Zone 6	8339
Warehouse_Ownership	25000	2	Company Owned	13578
approved_wh_govt_certificate	24092	5	C	5501

Table 6: Summary of Key Categorical Variables

Observations:

1. **Rural Warehouses & Logistics Challenges:**
   - A majority of warehouses (92%) are located in rural areas, which could lead to higher transportation costs, delays, and infrastructure constraints.
   - **Recommendation**: Consider optimizing transportation routes, investing in better road connectivity, or expanding urban warehouse locations to improve efficiency and reduce logistics expenses.
2. **Warehouse Capacity Distribution:**
   - Large and medium-capacity warehouses dominate, with few small-capacity warehouses.
   - **Recommendation**: If demand varies across regions, scaling small warehouses or upgrading medium ones could help balance storage utilization and prevent overstocking or shortages.
3. **Regional Warehouse Density:**
   - Zone 6 has the highest warehouse density, while Zone 1 has the lowest.
   - **Recommendation**: Assess warehouse placement strategy based on market demand and transportation feasibility. Consider expanding warehouses in underrepresented areas (like Zone 1) to ensure better market reach.
4. **Ownership Structure of Warehouses:**
   - Most warehouses (54%) are company-owned, which provides more control over operations.
   - **Recommendation**: Explore the benefits of leasing or partnering with third-party warehouses in strategic locations to reduce fixed costs and improve scalability.
5. **Warehouse Distribution Across Regions:**
   - North has the highest warehouse density, while East has the lowest.
   - **Recommendation**: Evaluate market demand in the Eastern region. If demand is high, establishing more warehouses in this area could reduce supply chain bottlenecks and improve service levels.

Bivariate Analysis:

Bivariate analysis means the analysis of the bivariate data. This is a single statistical analysis that is used to find out the relationship that exists between two value sets. The variables that are involved are X and Y.

- Univariate analysis is when only one variable is analyzed.
- Bivariate data analysis is when exactly two variables are analyzed.
- Multivariate analysis is when more than two variables get analyzed.

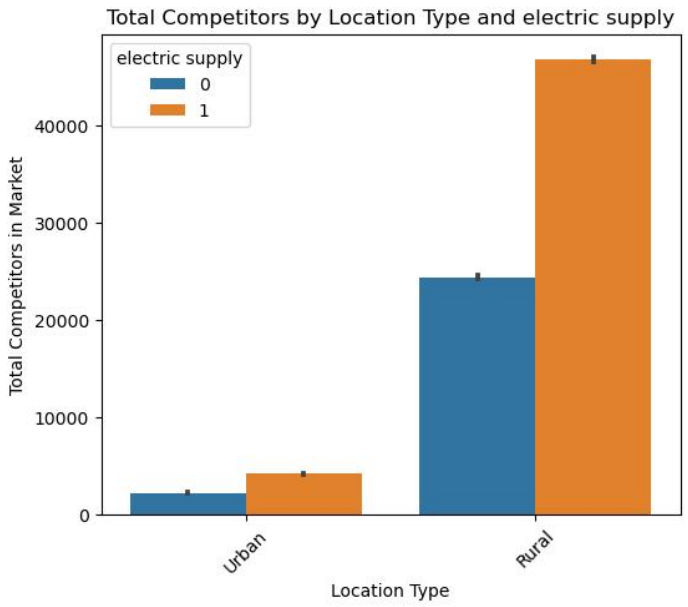

Figure 4 :Competitors by Location Type and electric supply

	Location_type	electric_supply	Competitor_in_mkt
0	Rural	0	24384
1	Rural	1	46792
2	Urban	0	2220
3	Urban	1	4209

Table 7 : Competitors by Location Type and electric supply

Impact of Electricity on Market Competition:

Rural areas with electricity have a significantly higher number of competitors (46,792) compared to rural areas without electricity (24,384). A similar trend is observed in urban areas, where electrified locations have 4,209 competitors, while non-electrified areas have only 2,220. This indicates that electricity access enhances operational efficiency and infrastructure, making these areas more attractive for businesses.

The lack of electricity in rural areas may be a key constraint for business expansion and competitor presence. To improve competitiveness, investing in electricity infrastructure—especially in rural areas—can enhance operational efficiency and attract more market players.

Additionally, in highly competitive electrified areas, businesses should focus on differentiation strategies such as service quality improvements and competitive pricing to maintain an edge. Conversely, non-electrified areas present a potential early-entry advantage, where companies can invest in alternative power sources (such as solar or generators) to gain a first-mover benefit and establish market dominance before competition intensifies.

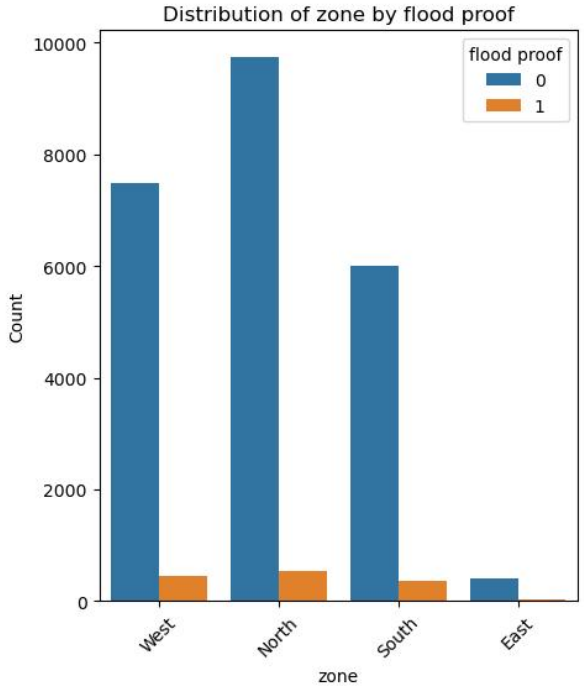

Figure 5 :Distribution of zone by flood proof

flood_proof	0	1
zone		
East	404	25
North	9737	541
South	6009	353
West	7484	447

Table 8 : Zone by flood proof

Flood-Proofing and Warehouse Vulnerability:

Across all zones, the majority of warehouses are not flood-proofed, posing significant operational risks. The North zone has the highest number of flood-proofed warehouses (541), while the East has the lowest (25), making it particularly vulnerable. Additionally, 9,737 warehouses in the North, 7,484 in the West, and 6,009 in the South remain at risk of flooding.

To mitigate these risks, priority should be given to flood-proofing efforts in the East and South zones, where protective measures are minimal. Identifying warehouses in flood-prone areas and investing in preventive infrastructure—such as drainage systems, elevated storage, and waterproof construction—will help reduce disruptions and losses during floods.

Further, implementing emergency preparedness measures—including backup power sources, contingency plans, and comprehensive insurance coverage—will ensure business continuity and financial protection against flood-related damages.

Figure 6 :Storage Issues Reported (Last 3 Months)

	govt_check_l3m	count	mean	min	max
0	1	550	17.141818	0	39
1	2	431	17.225058	0	39
2	3	438	17.221461	0	39
3	4	99	17.131313	0	39
4	5	250	16.764000	0	39
5	6	1224	17.540850	0	39
6	7	65	15.184615	0	35
7	8	276	17.003623	0	39
8	9	932	17.032189	0	39
9	10	899	16.846496	0	39
10	11	1160	17.150000	0	39
11	12	947	17.657867	0	39
12	13	429	17.069930	0	39
13	14	1429	17.261022	0	39
14	15	689	16.928882	0	39
15	16	201	17.263682	0	39
16	17	497	17.338028	0	39
17	18	217	17.138249	0	39
18	19	1604	17.230050	0	39
19	20	108	16.055556	0	37
20	21	649	16.935285	0	39
21	22	309	17.145631	0	39
22	23	1828	17.473742	0	39
23	24	628	17.335987	0	39
24	25	884	16.680995	0	39
25	26	2908	17.273384	0	39
26	27	1277	16.811276	0	39
27	28	1465	16.802048	0	39
28	29	901	16.964484	0	39
29	30	404	16.806931	0	39
30	31	362	15.914365	0	39
31	32	940	17.393617	0	39

Table 9 : Storage Issues Reported

Storage Issues:

Despite frequent government inspections, storage issues persist, indicating that regulatory oversight alone is not sufficient to resolve the problem. Warehouses with fewer government checks (e.g., 7 checks) report lower storage issues, while others with more frequent checks (e.g., 12 checks) experience higher storage issues . This suggests that internal inefficiencies in warehouse operations, rather than a lack of inspections, may be the root cause.

Notably, while some warehouses report zero storage issues, others consistently struggle, with storage issues peaking at 39. This disparity highlights the need for targeted interventions.

Analyze warehouses that consistently report high storage issues despite frequent inspections to pinpoint operational inefficiencies.Improve inventory management, space utilization, and material handling efficiency through structured audits.To improve storage efficiency adopt regular stock rotation, optimized shelving techniques, and enhanced climate control to minimize storage-related challenges, develop training programs focused on best practices in storage, damage

prevention, and inventory organization and Invest in automated tracking systems, real-time monitoring, and AI-driven inventory predictions to reduce inefficiencies and enhance storage management.

Multivariate Analysis:

Multivariate analysis is a statistical method that analyzes multiple variables simultaneously to find patterns and correlations. It's used to understand how different factors impact each other and to make predictions.

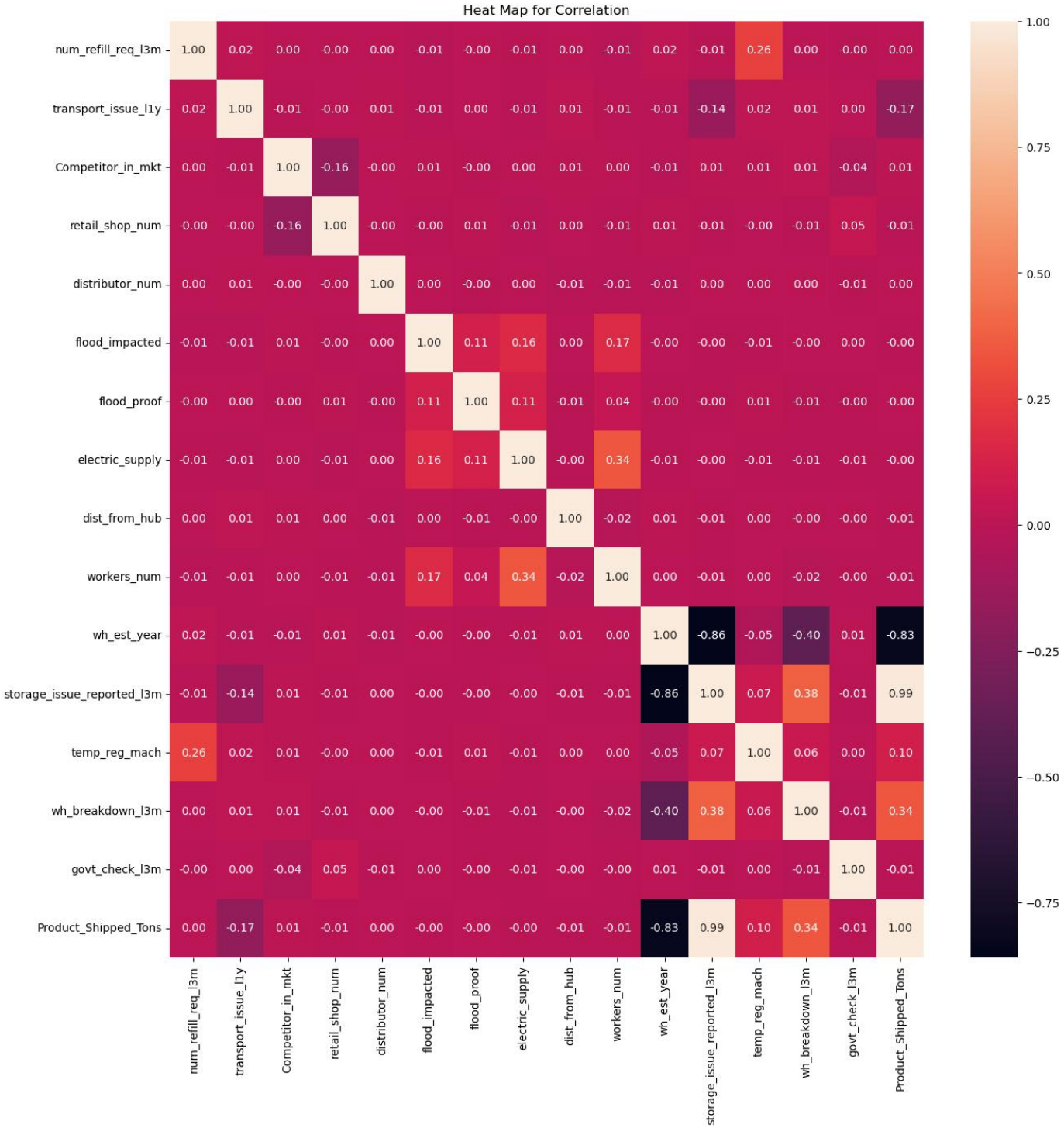

Figure 7 : Correlation HeatMap

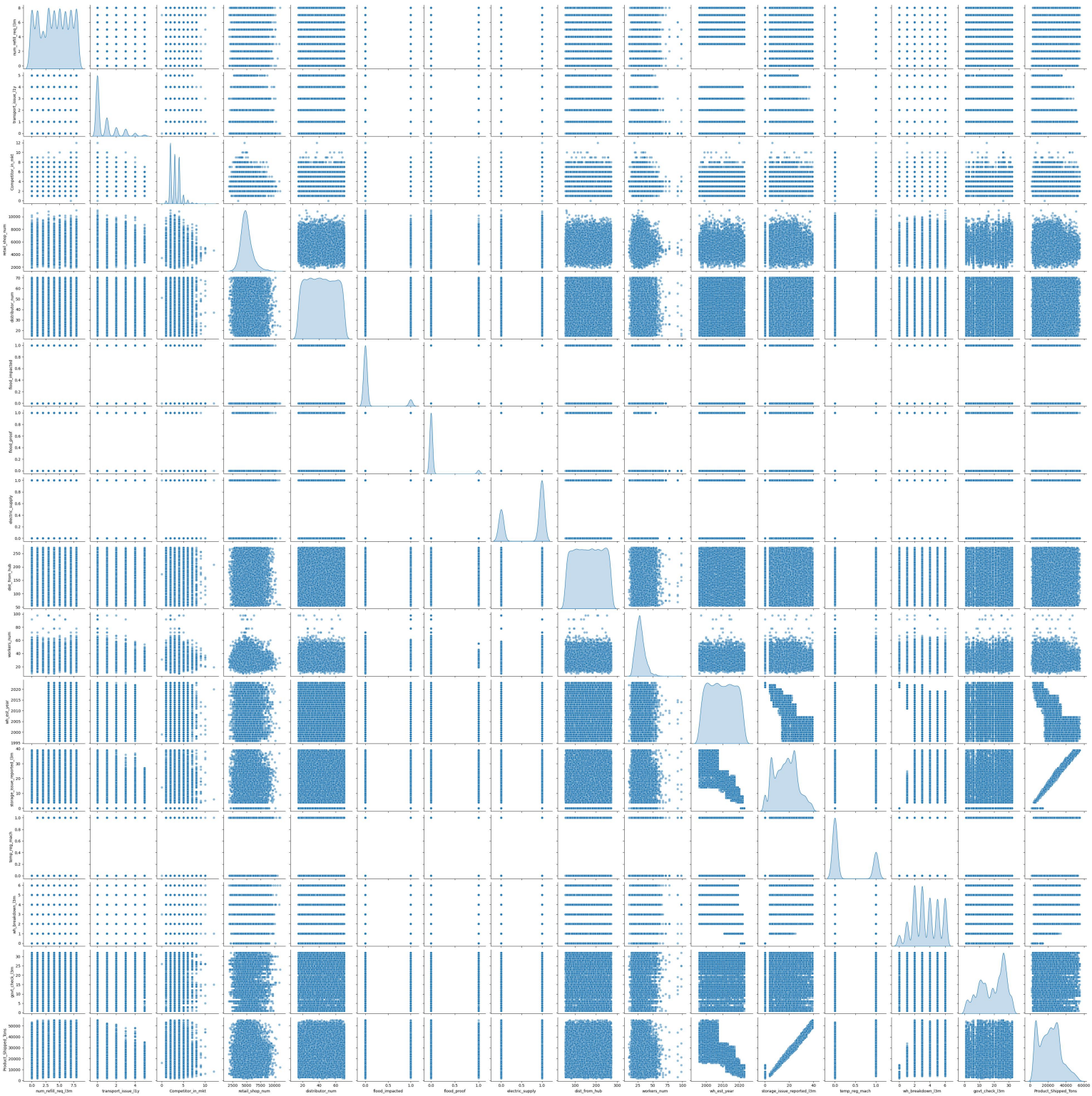

Figure 8 : Pairplot

Observations:

- There is a strong negative correlation between **storage issues** and **warehouse establishment year** (-0.86), indicating that older warehouses tend to experience more storage-related problems. This could be due to outdated infrastructure, poor space utilization, or inadequate inventory management systems.
- A high positive correlation (0.99) between **product shipped** and **storage issues** suggests that warehouses handling higher shipment volumes struggle with storage constraints. Efficient inventory management strategies, better shelving, and optimized storage layouts could help mitigate these challenges.
- Flood impact has a moderate correlation with electricity supply (0.16) and workforce (0.17), suggesting that flood-prone areas may still attract businesses due to better infrastructure. However, flood-proofing measures show only a weak correlation with storage issues, indicating that flood-proofing alone is not a strong determinant of storage efficiency.

To enhance warehouse efficiency and mitigate operational challenges, several strategic improvements can be implemented. Upgrading older warehouses with modern storage solutions can significantly reduce storage-related issues by optimizing space utilization and improving inventory management. In high shipment volume areas, enhancing warehouse infrastructure is crucial to effectively manage storage constraints and prevent bottlenecks. Additionally, improving preventive maintenance can help reduce warehouse breakdowns and minimize storage inefficiencies, ensuring smooth operations. For warehouses located in areas with low electricity supply, investing in alternative power solutions such as solar energy or backup generators can boost workforce availability and overall efficiency. Furthermore, optimizing inventory management strategies will help prevent overstocking issues, particularly in temperature-controlled warehouses, where proper stock rotation and forecasting are essential. Lastly, continuous monitoring of flood-prone zones and implementing proactive flood-proofing measures—especially in high-traffic warehouses—will reduce operational disruptions and financial risks. Implementing these strategies will enhance warehouse performance, minimize risks, and improve overall supply chain efficiency.

Removal Of Unwanted Variables

Certain columns might not be useful for analysis, such as unique identifiers or redundant information.These columns are unique for each entry and do not contribute to predictive analysis.

In our analysis, we dropped the columns '*Warehouse_ID*' and '*Warehouse_Manager_ID*'.The Warehouse ID and Warehouse Manager ID are unique identifiers assigned to each warehouse and its respective manager. These identifiers do not provide meaningful insights into operational efficiency, storage issues, or other key business metrics.Since they are categorical identifiers with no numerical significance, they do not contribute to trend analysis, correlation studies, or predictive modeling.Retaining these columns could lead to unnecessary complexity in data processing and visualization without adding value.These identifiers do not impact warehouse performance directly, so including them in calculations like correlations, summary statistics, or predictive models would be redundant.In machine learning models or statistical analysis, keeping unique identifiers can introduce noise, leading to overfitting or inaccurate insights.By dropping these columns, we ensure that our models focus on operational and performance-related variables rather than unique IDs.

In conclusion,Dropping '*Warehouse_ID*' and '*Warehouse_Manager_ID*' was a strategic data-cleaning decision aimed at improving data relevance, analytical accuracy, privacy compliance, and model efficiency. This step ensures that the analysis remains focused on actionable insights, helping businesses optimize warehouse operations and improve overall supply chain efficiency.

Missing Value treatment

Treating missing values in a regression model is crucial because leaving them unaddressed can lead to biased and unreliable results, as the model cannot accurately estimate relationships between variables when data is missing, potentially distorting the conclusions drawn from the analysis; essentially, missing data can introduce systemic errors and reduce the representativeness of your sample, impacting the validity of your model predictions.

Location_type	0
WH_capacity_size	0
zone	0
WH_regional_zone	0
num_refill_req_l3m	0
transport_issue_l1y	0
Competitor_in_mkt	0
retail_shop_num	0
Warehouse_Ownership	0
distributor_num	0
flood_impacted	0
flood_proof	0
electric_supply	0
dist_from_hub	0
workers_num	990
wh_est_year	11881
storage_issue_reported_l3m	0
temp_reg_mach	0
approved_wh_govt_certificate	908
wh_breakdown_l3m	0
govt_check_l3m	0
Product_Shipped_Tons	0

dtype: int64

Table 10 : List of Missing values.

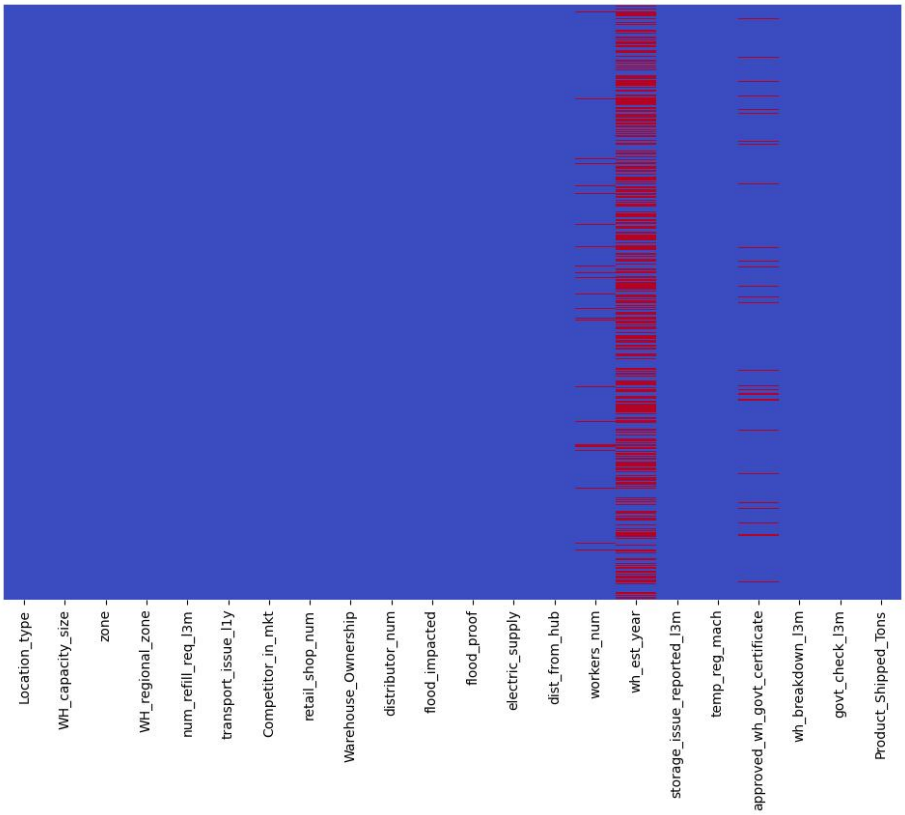

Figure 9 : Visual Representation of Missing values

The warehouse establishment year (wh_est_year) has nearly half of its data missing.This column was dropped because retaining it could introduce bias, and imputing such a large percentage could distort analysis.The missing data might be due to older warehouses lacking proper record-keeping.

The number of workers at each warehouse has a small percentage of missing values.Instead of dropping the column, we imputed missing values with the median to ensure data completeness while minimizing the effect of outliers.

Some warehouses are missing government certification records.Missing values were filled using the mode (most frequently occurring value), assuming that most warehouses follow regulatory compliance.Warehouses without certification might require further investigation to ensure compliance.

Outlier treatment:

Outlier treatment is crucial for regression models because outliers, which are data points significantly different from the rest of the data, can severely skew the regression line, leading to inaccurate model estimations and unreliable predictions if left unaddressed; essentially, a single outlier can disproportionately influence the model fit, compromising the overall analysis and insights derived from it.

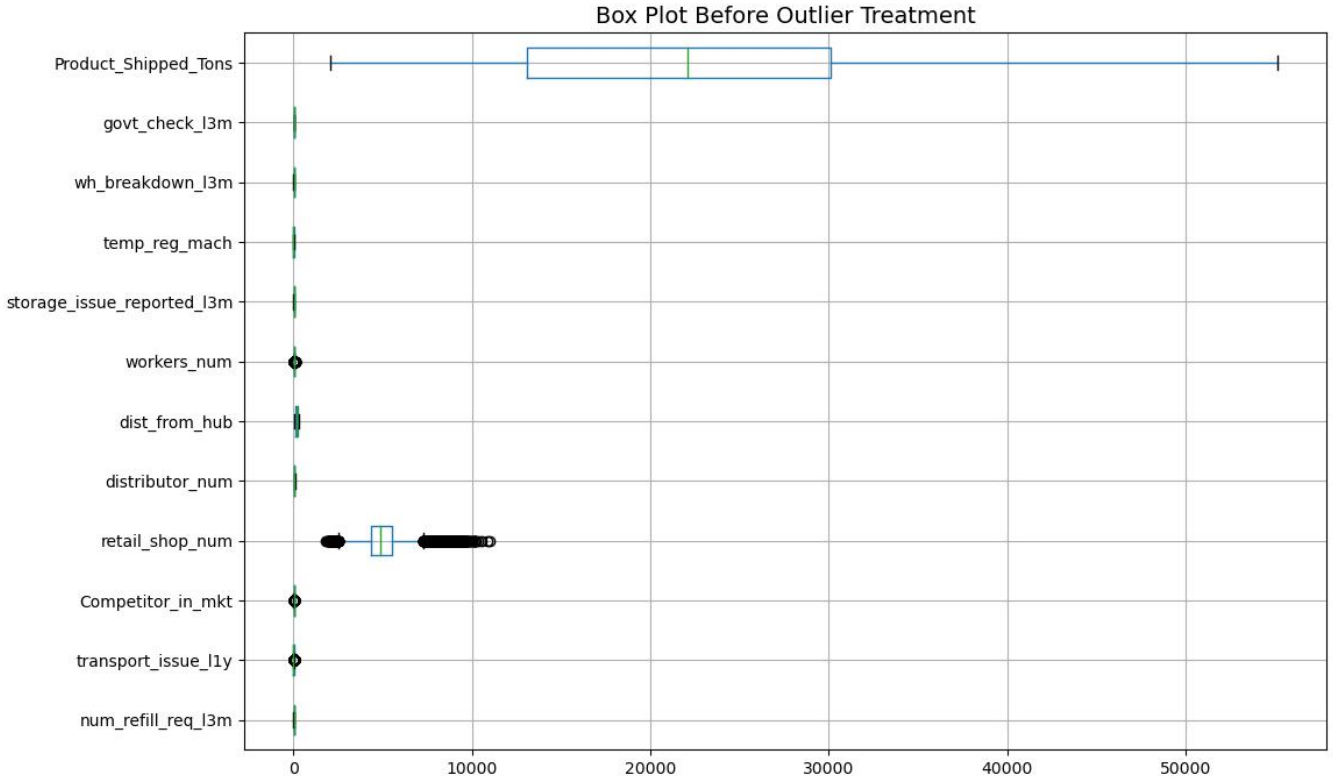

Figure 10 : Box Plot Before Outlier Treatment

We are using IQR method for Outlier detuction.The Interquartile Range (IQR) Method is a statistical technique used to detect and handle outliers in a dataset. It helps businesses identify unusual variations in key performance indicators, such as warehouse storage efficiency, shipment delays, or maintenance issues, which can impact decision-making.

The IQR method works by dividing the dataset into four equal parts (quartiles):

- Q1 (First Quartile - 25th percentile): The median of the lower half of the dataset.
- Q2 (Median - 50th percentile): The middle value of the dataset.
- Q3 (Third Quartile - 75th percentile): The median of the upper half of the dataset.
- **Interquartile Range (IQR):** The range between Q1 and Q3, calculated as:$IQR = Q3 - Q1$
- **Outlier Thresholds:**
    - **Lower Bound:** $Q1 - 1.5 \times IQR$
    - **Upper Bound:** $Q3 + 1.5 \times IQR$

    Any data point that falls below the lower bound or above the upper bound is considered an outlier.

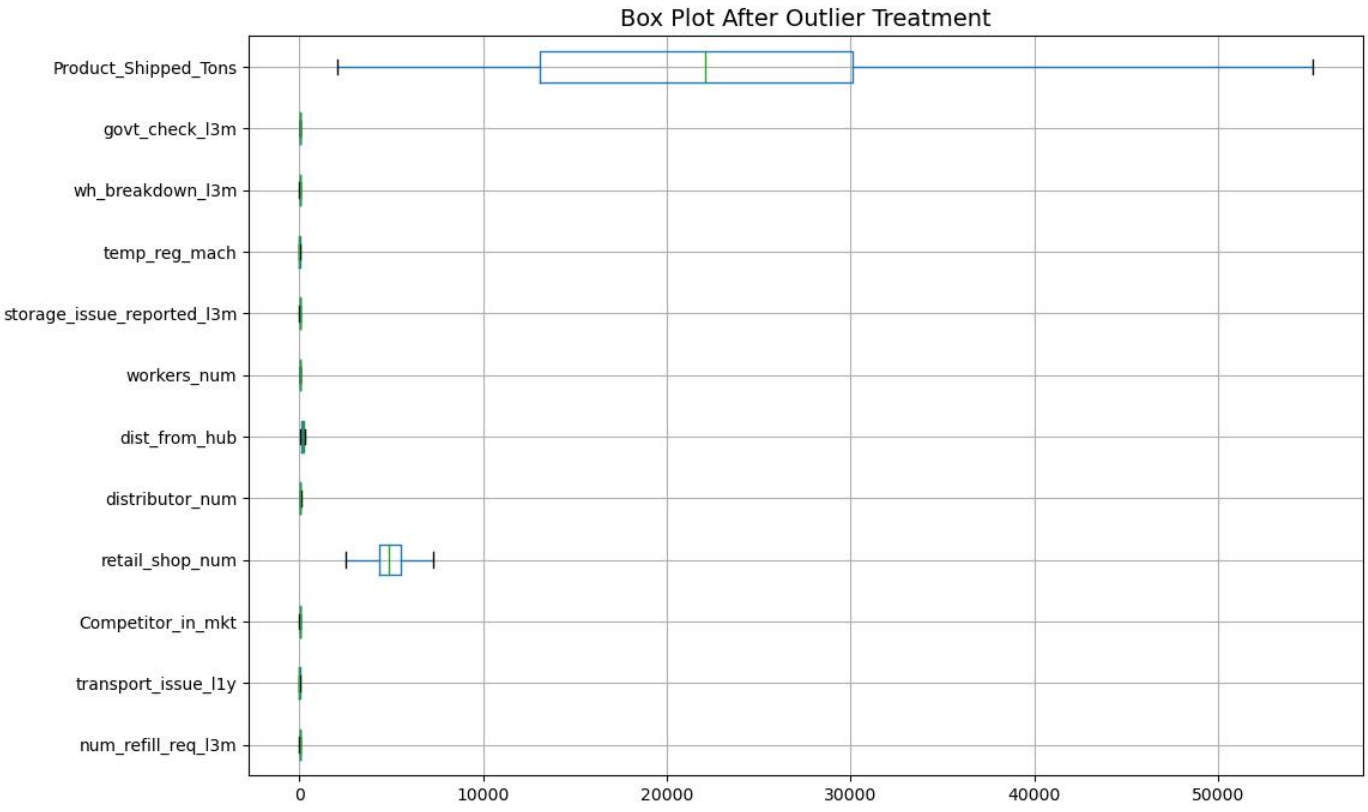

Figure 11 : Box Plot After Outlier Treatment

Variable transformation:

To ensure accurate and efficient analysis, the dataset undergoes several preprocessing steps, including scaling, encoding, and transformation. These techniques help standardize the data, making it more suitable for machine learning models and business insights.

Standardization of Continuous Variables:

In warehouse operations, different metrics, such as storage capacity, shipment volumes, and distances, exist on varying scales. For example, warehouse distance from hubs could be in kilometers, while shipment volumes are measured in tons. If left unstandardized, larger values could dominate smaller ones, leading to biased results in predictive models.

To address this, standardization is applied, which:

- Transforms numerical values so that they follow a common scale with a mean of 0 and a standard deviation of 1.
- Improves model performance by preventing certain variables from disproportionately influencing the analysis.
- Enhances comparability between different warehouse performance metrics, allowing fairer assessments.

Encoding Categorical Variables:

Warehouses are classified based on zones, regional zones, and ownership types, which are categorical variables. These need to be converted into a format suitable for data modeling.

One-Hot Encoding: This method creates new columns for each category while avoiding redundancy. For example:

- If there are three types of warehouse ownership (Rented,Company Owned), it creates two columns: Rented and Company Owned, with values as 1 or 0, indicating their presence.
- This transformation ensures that machine learning models can process categorical information effectively.

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	wo
0	-0.417807	0.374779	-0.980772	-0.317618	-1.146546	0	1	1	-1.156575	
1	-1.568750	-0.714377	0.803748	1.297843	0.285226	0	0	1	0.740827	
2	-1.185102	-0.714377	0.803748	-0.673514	1.343493	0	0	0	-0.040456	
3	1.116783	2.008512	-0.980772	1.073989	0.471979	0	0	0	-0.965240	
4	-0.417807	0.374779	-0.980772	-0.225807	-0.026028	1	0	1	-0.821739	

Table 11 : Datframe after transformation

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 31 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   num_refill_req_l3m              25000 non-null  float64
 1   transport_issue_l1y             25000 non-null  float64
 2   Competitor_in_mkt               25000 non-null  float64
 3   retail_shop_num                 25000 non-null  float64
 4   distributor_num                 25000 non-null  float64
 5   flood_impacted                  25000 non-null  int64  
 6   flood_proof                     25000 non-null  int64  
 7   electric_supply                 25000 non-null  int64  
 8   dist_from_hub                   25000 non-null  float64
 9   workers_num                     25000 non-null  float64
 10  storage_issue_reported_l3m      25000 non-null  float64
 11  temp_reg_mach                   25000 non-null  float64
 12  wh_breakdown_l3m                25000 non-null  float64
 13  govt_check_l3m                  25000 non-null  float64
 14  Product_Shipped_Tons            25000 non-null  float64
 15  zone_North                      25000 non-null  int8   
 16  zone_South                      25000 non-null  int8   
 17  zone_West                       25000 non-null  int8   
 18  WH_regional_zone_Zone 2         25000 non-null  int8   
 19  WH_regional_zone_Zone 3         25000 non-null  int8   
 20  WH_regional_zone_Zone 4         25000 non-null  int8   
 21  WH_regional_zone_Zone 5         25000 non-null  int8   
 22  WH_regional_zone_Zone 6         25000 non-null  int8   
 23  Warehouse_Ownership_Rented      25000 non-null  int8   
 24  Location_type_Urban             25000 non-null  int8   
 25  WH_capacity_size_Mid            25000 non-null  int8   
 26  WH_capacity_size_Small          25000 non-null  int8   
 27  approved_wh_govt_certificate_A+ 25000 non-null  int8   
 28  approved_wh_govt_certificate_B  25000 non-null  int8   
 29  approved_wh_govt_certificate_B+ 25000 non-null  int8   
 30  approved_wh_govt_certificate_C  25000 non-null  int8   
dtypes: float64(12), int64(3), int8(16)
memory usage: 3.2 MB
```

Table 12 : Datframe info after transformation

Business insights from EDA

Is the data unbalanced? If so, what can be done?

In classification, imbalance means one class significantly outweighs others, leading to biased model predictions. However, in regression, the concern is skewed or non-uniform distributions of the target variable rather than class imbalance.

Since we're working with regression, data imbalance isn't a direct issue, but target distribution matters. If your target variable is skewed or has high variance, you should apply transformations or outlier handling to improve model performance.

What is Balanced and Imbalanced Data?:

Imbalanced Data : Classification models attempt to categorize data into different buckets. In an imbalanced dataset, one bucket makes up a large portion of the training dataset (the majority class), while the other bucket is underrepresented in the dataset (the minority class). The problem with a model trained on imbalanced data is that the model learns that it can achieve high accuracy by consistently predicting the majority class, even if recognizing the minority class is equal or more important when applying the model to a real-world scenario.

Balanced Data : A balanced dataset is a data set where the classes or categories are represented in roughly equal proportions. This is important in machine learning for classification tasks, where it helps ensure the model is not biased towards any one class. In a balanced dataset, the number of samples from each class is roughly equal.Balanced datasets help achieve more accurate and reliable predictions. This is especially important when the cost of misclassification is high.

How to treat Imbalanced Data?:

Imbalanced data occurs when one category of the target variable significantly outweighs the others, leading to biased model predictions. This is especially relevant in classification problems where one class dominates, potentially reducing model accuracy for underrepresented categories.Below are strategies to handle data imbalance.

- Oversampling
- Undersampling
- Class Weight Adjustments
- Feature Engineering

Undersampling and Oversampling:

Oversampling : Oversampling involves increasing the number of instances in the minority class by duplicating existing samples or generating synthetic data. This ensures that the model gets sufficient exposure to the underrepresented class.

We can use SMOTE(Synthetic Minority Over-sampling Technique) for Oversampling. This creates synthetic samples by interpolating between existing data points in the minority class which reduces the risk of overfitting compared to simple duplication. This method works well when the minority class is not extremely small.

Few advantages of oversampling are this helps the model learn from the minority class and works well for small datasets where losing data (through undersampling) isn't an option.

Few disadvantages are it increases dataset size, leading to longer training times and if not done correctly, can lead to overfitting on the minority class.

Undersampling : Undersampling involves removing data points from the majority class to balance the dataset. This helps ensure that the model does not learn a bias toward the dominant class.

We can use Cluster-Based Undersampling randomly removing data, it groups similar majority class samples using clustering techniques (e.g., K-Means) and removes redundant ones. This retains the most informative samples from the majority class.

Few advantages of Undersampling are it reduces training time by making datasets smaller and ensures that the model is exposed to more balanced decision-making.

Few disadvantages are valuable data is lost, which may impact model performance and is not be effective when majority class data is highly diverse.

Best practise for unbalanced data is Start with undersampling if the dataset is large. Then use SMOTE if undersampling results in loss of important information and always evaluate results using precision, recall, F1-score, and AUC-ROC instead of accuracy alone.

Business insights using clustering:

We conducted K-Means clustering.K-Means is an unsupervised machine learning algorithm used for clustering similar data points into groups (clusters). It helps businesses identify patterns and segment data for better decision-making. The algorithm works by:

- **Selecting 'k' clusters** and randomly initializing centroids.
- **Assigning each data point** to the closest centroid (based on Euclidean distance).

- **Recalculating centroids** by taking the mean of all points assigned to each cluster.
- **Repeating steps 2 and 3** until convergence (centroids no longer change significantly)

In this analysis, we used K-Means clustering to segment warehouse operational data, helping businesses make strategic decisions based on warehouse performance and logistics efficiency.

We initialized K-Means clustering with k=2 to segment the dataset into two clusters.The inertia (Sum of Squared Errors, SSE) was calculated as 339070.27 and we extracted the cluster labels and stored them in our dataset.

To find the best 'k' value, we ran K-Means for k=1 to 10 and recorded the inertia values, higher inertia means data points are spread out (not well clustered).The elbow method helps find the optimal k by locating the point where the inertia stops decreasing significantly.

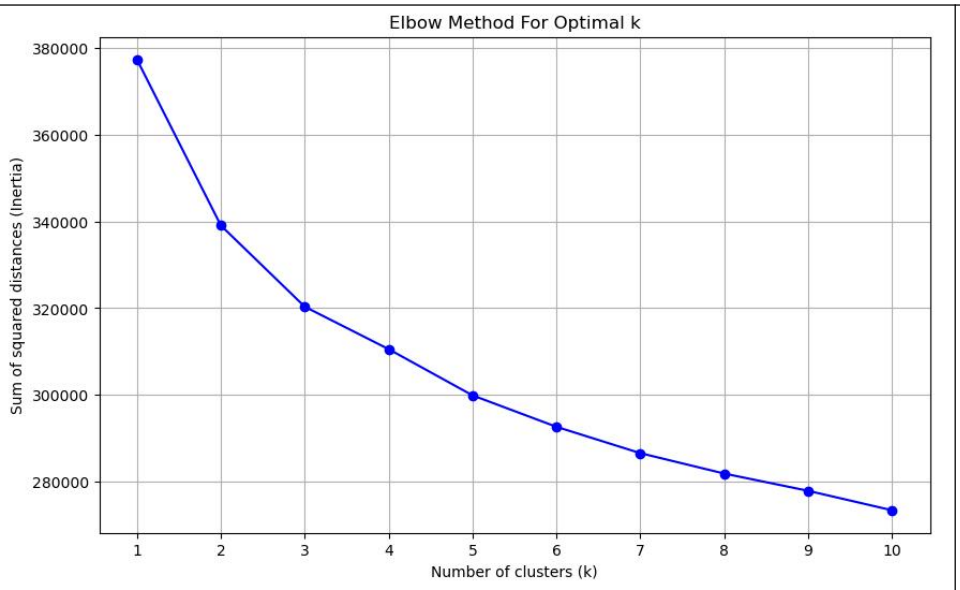

Figure 12:Elbow Method For Optimal k

```
[377350.59,
 339070.27,
 320335.77,
 310544.15,
 299886.48,
 292638.47,
 286559.84,
 281854.4,
 277877.13,
 273392.76]
```

Table 13:Within Sum Squares

Silhouette scores for each clusters are calculated and given below:

- For n_clusters=2, the silhouette score is 0.0949951632377869
- For n_clusters=3, the silhouette score is 0.08427410291857144
- For n_clusters=4, the silhouette score is 0.07982312994709234
- For n_clusters=5, the silhouette score is 0.07042472517199151
- For n_clusters=6, the silhouette score is 0.06250940397601544
- For n_clusters=7, the silhouette score is 0.057372113404488036
- For n_clusters=8, the silhouette score is 0.06118154150868958
- For n_clusters=9, the silhouette score is 0.05757868168989921

- For n_clusters=10, the silhouette score is 0.05270176990043396

We choose number of clusters as 3,which offered a reasonable balance between granularity and separation.

Cluster	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	dist_from_hub	workers_num	storage_issue_reported_l3m	temp_reg_mach	wh_breakdown_l3m	govt_check_l3m	Product_Shipped_Tons
0	0.41549080061391 24	-0.035555438	0.021976238152143827	-0.004400911	-0.002810778	-0.003916723	-0.003314099	0.41866952000 27657	1.515679068 5287067	0.296228773 42531406	0.005593960 351931496	0.482692792 7535891
1	-0.04130 73	0.230544185 19374692	-0.026849808	0.017139478 602005848	-0.00439313	0.023157902 439019797	0.012784674 395271041	-1.059124551	-0.276079094	-0.679284199	0.019570513 732234656	-1.043328695
2	-0.20 4131 316	-0.166097977	0.008947952 34432606	-0.011323812	0.005198417 480982213	-0.016483125	-0.008428367	0.61422655746 59252	-0.659770278	0.377790900 04020091	-0.019119857	0.564106930 1013012

Table 14 : Cluster Summary

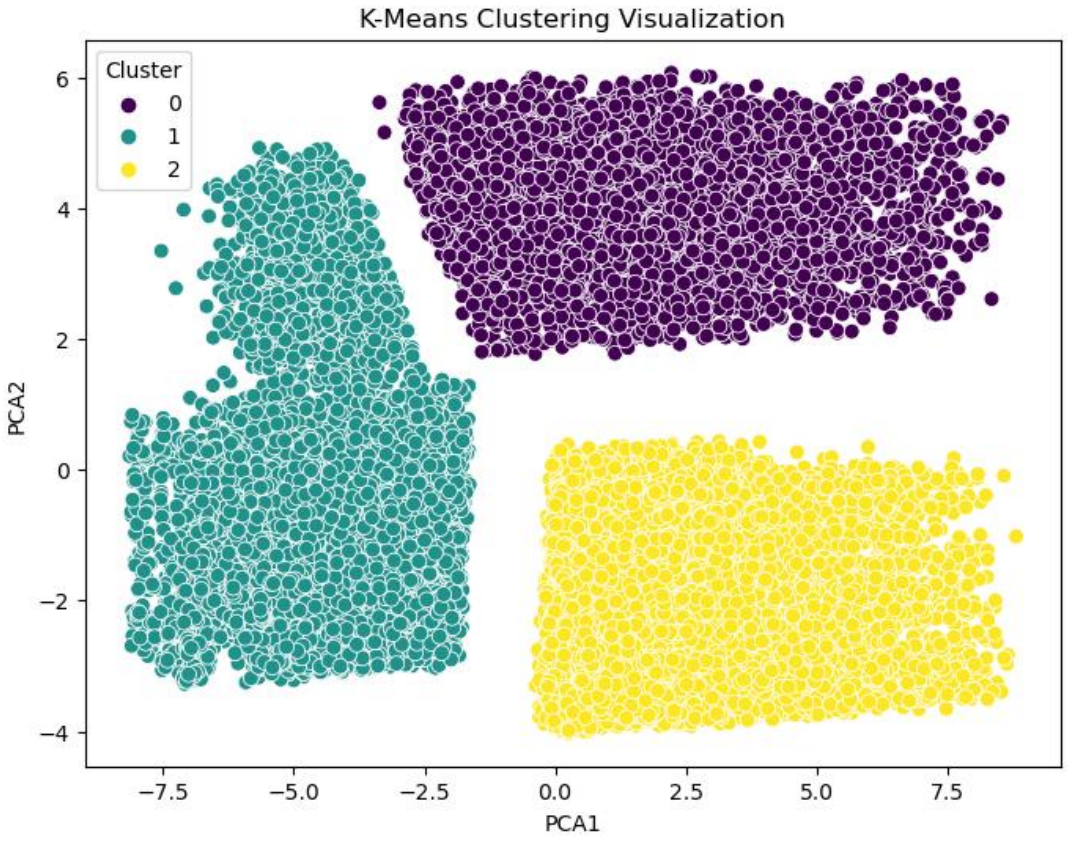

Figure 13:K-Means Clustering Visualization

The table represents the average values of key warehouse performance indicators across three clusters. These variables were chosen because they directly impact warehouse efficiency, supply chain performance, and operational risks.

1. **Cluster 0** – High-Performing Warehouses (Efficient but Storage Issues)
    - Frequent refills and high shipment volume indicate strong demand.
    - Temperature regulation machines are widely available, suggesting cold storage or perishable goods.
    - Higher storage issues reported, indicating a need for better space management.
    - Action: Improve warehouse capacity planning, optimize storage layout, and ensure smooth logistics.

2. **Cluster 1** – Problematic Warehouses (Logistics Challenges & Low Shipments)
    - Frequent transport issues and high government checks suggest regulatory compliance challenges.
    - Fewer storage issues reported might indicate unused storage capacity.
    - Low shipment volume, meaning these warehouses are underutilized.
    - Action: Improve transportation infrastructure, address compliance issues, and analyze whether to shut down or repurpose these warehouses.

3. **Cluster 2** – Medium Performance Warehouses (Breakdowns & Moderate Shipments)
    - Moderate refill demand and medium shipment volume suggest fluctuating business activity.
    - High warehouse breakdown frequency, meaning poor maintenance and aging infrastructure.
    - Fewer government inspections, which could indicate compliance risks.
    - Action: Invest in preventive maintenance, enforce regular quality checks, and ensure compliance with government regulations.

Other Business Insights:

Warehouse Clustering & Business Segmentation : Using K-Means Clustering, we segmented warehouses into three clusters based on key operational factors such as the number of refill requests, transport issues, competitor presence, retail shop numbers, distributor count, and storage-related inefficiencies. The insights from each cluster are:

- Warehouses in cluster 0 have a higher number of refill requests and report increased activity related to temperature regulation and storage management.These warehouses are actively used and require better preventive maintenance and expansion to avoid bottlenecks.Investing in automated inventory tracking and real-time monitoring solutions will prevent delays and optimize resource allocation.
- Cluster 1 has higher transport-related issues and negative government inspection results, along with higher breakdown occurrences.These warehouses face operational inefficiencies, poor logistics support, and compliance issues.Enhancing transportation routes, improving warehouse maintenance schedules, and ensuring compliance with government safety regulations will improve efficiency.
- Warehouses in cluster 2 have fewer refill requests, lower competitor impact, and fewer workers, indicating underutilization.These warehouses may be operating below capacity and could be optimized for better resource allocation or repurposing.Conducting market analysis to identify demand gaps and optimizing inventory flow can enhance warehouse efficiency.

Infrastructure & Maintenance Priorities : Warehouses in flood-affected areas face operational risks. Implementing flood-proofing measures and investing in drainage solutions will improve long-term viability.Warehouses with limited electricity supply impact workforce efficiency. Investing in alternative power solutions (solar, backup generators) can enhance workforce productivity and reduce downtime.

Inventory Management Optimization : Some temperature-controlled warehouses face overstocking problems, which can lead to spoilage and increased costs.Implementing demand forecasting models using historical shipping trends can optimize stock levels and reduce wastage.

Transport & Logistics Optimization : Warehouses with higher transport issues may suffer from poor road connectivity, inadequate vehicle availability, or inefficient routing.Route optimization algorithms and AI-powered logistics planning can reduce transportation delays and improve delivery efficiency.

Predictive Maintenance & Preventive Strategies : Warehouses reporting frequent breakdowns tend to have higher inefficiencies in storage operations.Implementing IoT-based predictive maintenance can help detect early signs of equipment failure, reducing downtime and increasing operational efficiency.

To enhance operational efficiency, underperforming warehouses should be upgraded with modern storage solutions and improved infrastructure. Investing in automation for inventory management and predictive maintenance can help minimize downtime and optimize resource utilization. Additionally, transport logistics should be improved by optimizing delivery routes and strategically selecting warehouse locations. Ensuring compliance with government regulations is crucial, particularly for warehouses with frequent inspection failures, to avoid penalties and operational disruptions. Implementing AI-driven demand forecasting will enable better stock level management, reducing waste and improving supply chain efficiency. Lastly, allocating resources based on warehouse cluster segmentation will help optimize costs and enhance overall performance.

Project Notes - 2

MODEL BUILDING

Overview of Models:

To enhance supply chain efficiency and optimize decision-making, various machine learning models were developed and tested. The models included:

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. Gradient Boosting Regression
5. AdaBoost Regression

These models were trained on historical supply chain data to predict key performance metrics.

Model Performance Evaluation:

Each model was tested using standard performance metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors.
- Root Mean Squared Error (RMSE): Captures large deviations by penalizing bigger errors more heavily.
- R^2 Score: Indicates how well the model explains variance in the target variable.

Initially we copied all predictor values into X dataframe and target (*Product_Shipped_Tons*) into y dataframe. Next we split X an y into training and test set in 75:25 ratio.

Linear Regression:

Linear regression is a machine learning algorithm that uses a linear equation to predict the value of a dependent variable based on an independent variable. It's a supervised learning algorithm that uses labeled datasets to learn and make predictions.

- Linear regression models the relationship between the variables as a linear equation.
- It uses the least squares method to find the best fit line that minimizes the distance between the line and the actual data points.

- The line of best fit is used to predict the value of the dependent variable.

Assumptions of Linear Regression:

- Linear regression assumes a linear relationship between the variables.
- It assumes that the differences between the predicted and observed values are the same for all independent variables.

Linear Regression Results:

	MAE	RMSE	R^2 Score
0	0.018673	0.025196	0.986485

Table 15 : Linear Regression Results

The model demonstrates strong predictive performance, with an average deviation of 0.0187 units from actual values. The Root Mean Squared Error (RMSE), which places greater emphasis on larger errors, is 0.025196, indicating high accuracy. Additionally, the R^2 score of 0.9856 suggests that 98.65% of the variance in the target variable is effectively explained by the model's features, confirming a strong fit to the data.

Given the high explanatory power and low error rates, the model is highly reliable for forecasting. The minimal deviation between predicted and actual values further supports its accuracy and consistency. If the business requires a simple yet interpretable model, Linear Regression serves as a robust and effective choice for data-driven decision-making.

Decision Tree:

Decision tree regression is a machine learning technique that predicts continuous numerical values using a tree-like model. It's a supervised learning algorithm that's used to model data and make predictions.

- Decision tree regression splits data into smaller subsets based on certain criteria.
- It predicts the average value of the target variable within each subset.

Decision Tree Regression Results:

	MAE	RMSE	R^2 Score
0	0.01648	0.024487	0.987236

Table 16:Decision Tree Regression Results

The Decision Tree model demonstrates a slightly lower average prediction error compared to Linear Regression (0.0187), along with a lower RMSE (0.0252), indicating improved accuracy. The R^2 score of 0.9872 suggests that 98.72% of the variance in the target variable is explained by the model, marking a marginal improvement over Linear Regression (98.65%), which translates to a better fit to the data.

While the model exhibits better predictive power, this comes at the expense of interpretability. Decision Trees are prone to overfitting, particularly if not pruned or properly tuned, which may affect generalizability in real-world applications. Careful hyperparameter tuning and pruning techniques are recommended to enhance model stability and reliability for business use.

Random Forest:

A Random forest regression model combines multiple decision trees to create a single model. Each tree in the forest builds from a different subset of the data and makes its own independent prediction. The final prediction for input is based on the average or weighted average of all the individual trees' predictions.

- Random forest regression builds many decision trees during training.
- Each tree is trained on a different subset of the data.
- The predictions from each tree are averaged to produce a single result.

Random forest Regression Results:

Random Forest Regression Results:

	MAE	RMSE	R^2 Score
0	0.013174	0.017566	0.993431

Table 17: Random Forest Results

The Random Forest model demonstrates superior accuracy and reliability compared to Decision Tree and Linear Regression. With a Mean Absolute Error (MAE) of 0.0132, it produces more precise predictions than Decision Tree (0.0165) and Linear Regression (0.0187). The Root Mean Squared Error (RMSE) of 0.0176 marks a significant improvement over Decision Tree (0.0245) and Linear Regression (0.0252), indicating reduced prediction errors. Additionally, the R^2 Score of 0.9934 confirms that the model explains 99.34% of the variance, making it the best-performing model in terms of predictive accuracy.

Unlike Decision Trees, which are prone to overfitting, Random Forest mitigates this issue by averaging multiple trees, leading to improved generalization. Its ability to handle non-linear relationships makes it a robust and versatile choice for business applications that demand high accuracy and stability in forecasting and decision-making.

Gradient Boosting:

Gradient Boosting Regression is a machine learning technique used for regression tasks, where it builds a predictive model by sequentially adding multiple "weak learner" decision trees, with each new tree focusing on correcting the errors made by the previous trees, resulting in a more accurate overall prediction; essentially, it's an ensemble method that iteratively improves upon the previous predictions by minimizing the residuals (errors) at each step.

1. The algorithm starts with a simple initial prediction, often the average of the target variable.

2. For each data point, the difference between the actual target value and the initial prediction is calculated, creating the "residuals".
3. A decision tree is trained to predict these residuals.
4. The prediction from the new decision tree is added to the previous prediction, creating a more refined prediction.
5. Steps 2-4 are repeated, building a series of decision trees, each focusing on correcting the errors from the previous iteration.

Gradient Boosting Regression Results:

Gradient Boosting Regression Results:

	MAE	RMSE	R^2 Score
0	0.012956	0.017047	0.993814

Table 18 : Gradient Boosting Results

The Gradient Boosting model outperforms Random Forest in terms of accuracy and predictive power. With a Mean Absolute Error (MAE) of 0.01296, it achieves slightly lower prediction errors than Random Forest (0.0132). The Root Mean Squared Error (RMSE) of 0.01705 is also a marginal improvement over Random Forest (0.01757), indicating better overall accuracy. Additionally, the R^2 Score of 0.9938 confirms that the model explains 99.38% of the variance, making it the most powerful model so far in terms of predictive performance.

Gradient Boosting proves to be more accurate than Random Forest, making it a strong candidate for production deployment. It is particularly effective at capturing complex, non-linear relationships and adapting to various data patterns. However, it might require hyperparameter tuning to prevent overfitting, though it generally performs well even with default settings.

AdaBoost:

AdaBoost regression is a machine learning technique that uses the AdaBoost (Adaptive Boosting) algorithm to perform regression tasks, where it combines multiple "weak" regression models (like decision trees) into a single "strong" regression model by iteratively adjusting weights on data points based on their previous prediction errors, giving more importance to difficult-to-predict instances in subsequent iterations, ultimately creating a more accurate prediction overall; essentially, it's an ensemble learning method for regression problems.

1) Each data point is assigned an equal weight initially.
2) Train a simple regression model (like a decision stump) on the current weighted data.
3) Adjust the weights of data points based on the errors made by the current weak learner.
4) Repeat steps 2 and 3, adding new weak learners to the ensemble, until a stopping criterion is met.
5) The final prediction is a weighted average of the predictions made by all the weak learners.

Adaboost Regression Results:

	MAE	RMSE	R^2 Score
0	0.026028	0.032538	0.977462

Table 19: Adaboost Results

The AdaBoost model demonstrates weaker performance compared to other ensemble methods, particularly Gradient Boosting and Random Forest. With a Mean Absolute Error (MAE) of 0.02603, it has higher prediction errors than both Gradient Boosting (0.01296) and Random Forest (0.0132). The Root Mean Squared Error (RMSE) of 0.03254 is the highest among all ensemble models, indicating less accurate predictions. Additionally, the R^2 Score of 0.9775, while still high, is lower than Gradient Boosting (0.9938) and Random Forest (0.9934), meaning it explains less variance in the data.

Overall, AdaBoost's performance is weaker compared to other ensemble models, suggesting that it may not be the best choice for this dataset. Since AdaBoost works best with simple base models, it might require further hyperparameter tuning or a stronger base estimator to achieve better accuracy.

Model Tuning

Model tuning is the process of optimizing machine learning models by adjusting hyperparameters to improve performance. Unlike model parameters (learned from data), hyperparameters are set before training and influence how the model learns.

Model Tuning is done so:

- Ensure the model generalizes well to unseen data.
- Minimize prediction mistakes in demand forecasting.
- Help management make data-driven inventory, marketing, and logistics decisions.
- Prevent the model from memorizing the training data or performing poorly on test data.

We initially trained Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost models and based on the initial results, Random Forest, Gradient Boosting, and AdaBoost were chosen for tuning.

We fine-tuned the models using GridSearchCV, which performs exhaustive searches over multiple hyperparameter combinations.

After tuning, we re-evaluated the models using MAE, RMSE, and R^2 Score to measure improvements after which model was selected for demand forecasting and business insights.

Gradient Boosting Tuning:

Hyperparameters tuned are Number of trees(n_estimators), *learning_rate*: control contribution of each tree and *max_depth:* limits complexity of trees.

Best parameters for Gradient Boosting: 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100.

Random Forest Tuning:

Hyperparameters tuned are *n_estimators* (Number of decision trees),*max_depth* (Controls depth of trees) and *min_samples_split* (Minimum samples to split a node).

Best parameters for Random Forest: 'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200

AdaBoost Tuning:

Hyperparameters tuned are *n_estimators* (Number of weak learners) and *learning_rate* (Influences the contribution of each learner).

Best parameters for AdaBoost: 'learning_rate': 0.1, 'n_estimators':

100.

	MAE	RMSE	R² Score
Tuned Random Forest	0.012772	0.016985	0.993859
Tuned Gradient Boosting	0.012571	0.016687	0.994072
Tuned AdaBoost	0.024818	0.031345	0.979085

Table 20: Results after Tuning

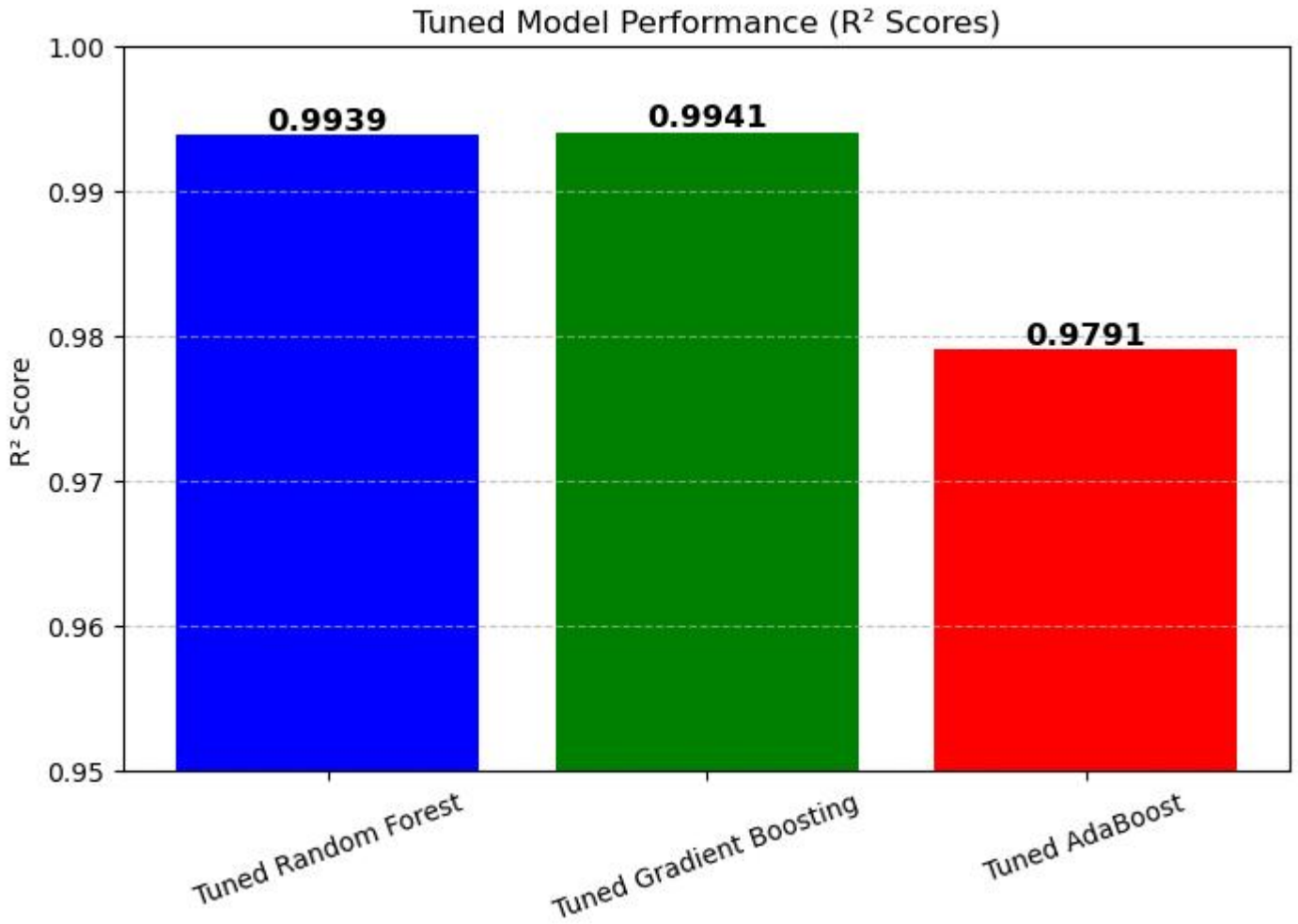

Figure 14:Tuned Model Performance (R² Scores)

Inference:

From the results, we see that Tuned Gradient Boosting Regressor achieved the best performance across all three metrics:

- Lowest Mean Absolute Error (MAE) :0.012571
- Lowest Root Mean Squared Error (RMSE) : 0.016687
- Highest R^2 Score : 0.994072

This indicates that Tuned Gradient Boosting makes the most accurate predictions with the least error. It explains 99.41% of the variance in the target variable, making it the most reliable model for business decision-making.

Business Implications of the Best Model:

Better Demand Forecasting : More accurate predictions allow efficient inventory planning, reducing overstocking or shortages.

Optimized Logistics & Supply Chain : Improved accuracy in demand prediction enables better warehouse stocking and delivery scheduling, reducing costs and improving efficiency.

Targeted Marketing Campaigns : Insights from the model can identify high-demand regions, helping management strategically allocate advertising budgets to maximize ROI.

Cost Savings & Revenue Growth : With precise demand forecasting, businesses can minimize wastage, operational inefficiencies, and unnecessary storage costs, ultimately boosting profitability.

Gradient Boosting works by iteratively improving weak models and minimizing prediction errors, making it well-suited for complex data patterns. It outperformed Random Forest & AdaBoost because:

- Handles non-linearity better than Random Forest.
- Less sensitive to noise than AdaBoost, which tends to overfit.
- Fine-tuned hyperparameters improved its generalization on unseen data.

Adopt Tuned Gradient Boosting for demand forecasting & inventory management,Use model insights to plan targeted marketing campaigns in high-demand areas and Integrate predictive analytics into supply chain decision-making to improve efficiency.

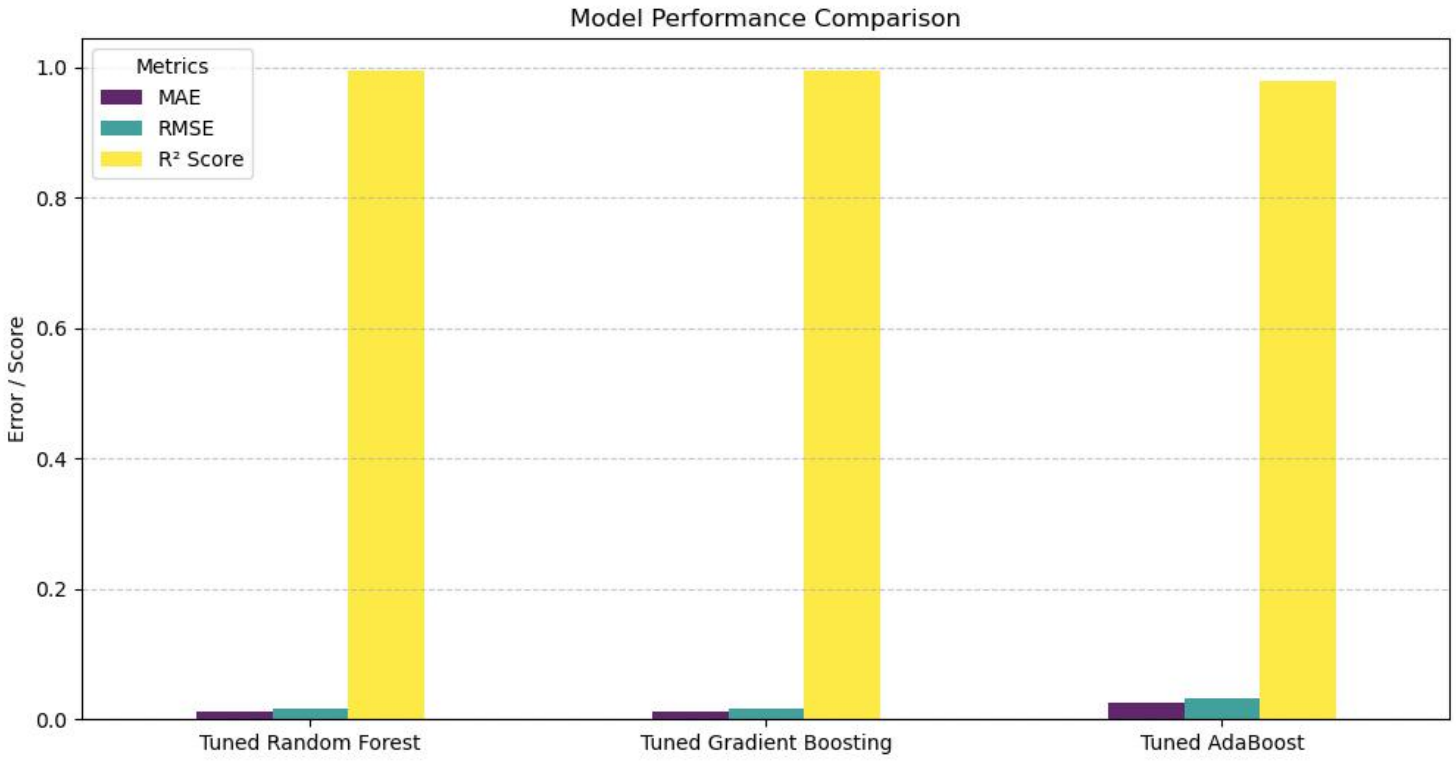

Figure 15:Model Performance Comparison

BUSINESS IMPLICATIONS & OPTIMIZATION STRATEGY BASED ON MODEL INSIGHTS

Determining the Optimum Weight of Products to Be Shipped to Warehouses:

- Based on the regression analysis, Gradient Boosting Regression emerged as the most optimal model with the lowest error metrics (MAE: 0.012571, RMSE: 0.016687, R^2: 0.994072).
- This model can be used to predict ideal shipment quantities based on historical demand patterns, minimizing overstocking and stockouts.
- Warehouses in Cluster 0 (high refill requests, storage issues) should receive frequent, smaller shipments to prevent bottlenecks.
- Warehouses in Cluster 2 (underutilized) should receive minimal shipments and be considered for repurposing or alternate storage strategies.
- Dynamic weight allocation strategy based on predicted demand patterns will improve efficiency and reduce waste.

Demand Pattern Analysis for Targeted Advertisement Campaigns:

- Using K-Means Clustering, we identified distinct warehouse clusters based on operational characteristics.
- High-demand areas (Cluster 0) indicate strong customer demand, suggesting that these regions can benefit from increased marketing and targeted advertisements.
- Heatmap analysis using warehouse location data can help pinpoint the most active regions for product demand.
- Competitive intelligence: Areas with high demand but lower competitor presence should be prioritized for brand awareness campaigns.

Strategic Next Steps:

- The company has provided limited data in the first phase; further analysis using their full 360-degree data lake will enhance forecasting accuracy.
- The Gradient Boosting model's performance indicates that machine learning-based inventory planning can significantly enhance operational efficiency.
- Future Steps:
  - Implement AI-driven demand forecasting at a granular, regional level.
  - Integrate real-time IoT sensors for warehouse monitoring.
  - Develop an automated dashboard to visualize demand trends dynamically.