Capstone Project

Supply Chain Project

DSBA Course

Rishabh Gupta

Made by –

Rishabh Gupta

Contents

Rishabh Gupta

Objective

A FMCG company has entered into the instant noodles business two years back. Their higher management has notices that there is a miss match in the demand and supply. Where the demand is high, supply is pretty low and where the demand is low, supply is pretty high. In both the ways it is an inventory cost loss to the company; hence, the higher management wants to optimize the supply quantity in each and every warehouse in entire country.

Problem Statement: 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 to the warehouse. Also try to analysis the demand pattern in different pockets of the country so management can drive the advertisement campaign particular in those pockets.

This is the first phase of the agreement; hence, company has shared very limited information. Once you are able to showcase a tangible impact with this much of information then company will open the 360-degree data lake for your consulting company to build a more robust model.

Target variable: product_wg_ton
Dataset: Data.csv

Rishabh Gupta

Problem Statement Analysis

Current Challenge:
The FMCG company's instant noodles business suffers from demand-supply mismatch, causing high inventory costs and lost sales due to stockouts, disrupting resources and market potential.

Why Align Supply with Demand?

- **Cut Inventory Costs:** Accurate forecasting minimizes excess stock and related expenses.
- **Boost Customer Satisfaction:** Optimal stock ensures availability, reducing waste and improving loyalty.
- **Improve Efficiency:** Balanced inventory streamlines logistics and productivity.
- **Enable Smarter Decisions:** Predictive insights guide production and marketing strategies.
- **Gain Competitive Advantage:** Agile supply chain adapts faster to market changes.

Project Impact:

- Develop a demand forecasting model to predict shipment weights per warehouse.
- Reveal regional demand patterns for targeted marketing.
- Lower operational costs via optimized inventory.
- Build a data-driven supply chain foundation for future growth and deeper optimization.

Next Steps:
By leveraging key variables, we will develop a model to determine the optimal *product_wg_ton* for each

warehouse, enhancing supply chain efficiency and cost-effectiveness. Additionally, analysing location-based demand drivers will refine marketing strategies for maximum impact.

Data Dictionary –

Sr. No	Column Name	Description
1	Ware_house_ID	Product warehouse ID
2	WH_Manager_ID	Employee ID of warehouse manager
3	Location_type	Location of warehouse like in city or village
4	WH_capacity_size	Storage capacity size of the warehouse
5	zone	Zone of the warehouse
6	WH_regional_zone	Regional zone of the warehouse under each zone
7	num_refill_req_l3m	Number of times refilling has been done in last 3 months
8	transport_issue_l1y	Any transport issue like accident or goods stolen reported in last one year
9	Competitor_in_mkt	Number of instant noodles competitor in the market
10	retail_shop_num	Number of retails shop who sell the product under the warehouse area
11	wh_owner_type	Company is owning the warehouse or they have get the warehouse on rent
12	distributor_num	Number of distributer works in between warehouse and retail shops
13	flood_impacted	Warehouse is in the Flood impacted area indicator
14	flood_proof	Warehouse is flood proof indicators. Like storage is at some height not directly on the ground
15	electric_supply	Warehouse have electric back up like generator, so they can run the warehouse in load shedding
16	dist_from_hub	Distance between warehouse to the production hub in Kms
17	workers_num	Number of workers working in the warehouse
18	wh_est_year	Warehouse established year
19	storage_issue_reported_l3m	Warehouse reported storage issue to corporate office in last 3 months. Like rat, fungus because of moisture etc

Rishabh Gupta

20	temp_reg_mach	Warehouse have temperature regulating machine indicator
21	approved_wh_govt_certificate	What kind of standard certificate has been issued to the warehouse from government regulatory body
22	wh_breakdown_l3m	Number of time warehouse face a breakdown in last 3 months. Like strike from worker, flood, or electrical failure
23	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
24	product_wg_ton	Product has been shipped in last 3 months. Weight is in tons

Data Overview –

The Dataset has 25000 number of rows with 24 columns.

	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_
0	WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	2	
1	WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	4	
2	WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	4	
3	WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	2	
4	WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	2	

TABLE 1 - TOP 5 ROWS OF DATASET

Rishabh Gupta

```
<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 DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe there are 16 numerical and 8 object type variables in dataset.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	work
count	25000.00	25000.00	25000.00	25000.00	25000.00	25000.0	25000.00	25000.00	25000.00	
mean	4.09	0.77	3.10	4985.71	42.42	0.1	0.05	0.66	163.54	
std	2.61	1.20	1.14	1052.83	16.06	0.3	0.23	0.47	62.72	
min	0.00	0.00	0.00	1821.00	15.00	0.0	0.00	0.00	55.00	
25%	2.00	0.00	2.00	4313.00	29.00	0.0	0.00	0.00	109.00	
50%	4.00	0.00	3.00	4859.00	42.00	0.0	0.00	1.00	164.00	
75%	6.00	1.00	4.00	5500.00	56.00	0.0	0.00	1.00	218.00	
max	8.00	5.00	12.00	11008.00	70.00	1.0	1.00	1.00	271.00	

TABLE 3 - STATISTICAL SUMMARY OF DATASET

Rishabh Gupta

Observations-

- The dataset mostly has 25,000 entries, except for wh_est_year with only ~13,000, indicating missing data.
- The target variable product_wg_ton shows high variability (mean ~22,100 tons, SD ~11,600) with a wide range.
- num_refill_req_13m averages 4 refills in 3 months, but some warehouses require up to 8.
- transport_issue_lly has a mean of 0.77, indicating moderate transport-related risks across sites.
- Competition varies widely (Competitor_in_mkt mean = 3.1, max = 12), reflecting diverse market pressures.
- retail_shop_num has high variation (mean ~4865, max >11,000), indicating large differences in retail reach.
- distributor_num averages 42, but ranges widely, affecting product flow.
- Flood vulnerability is low (flood_impacted = 0.1, flood_proof = 0.05), but still relevant for risk planning.
- Around 66% of warehouses have electric backups, ensuring resilience during power cuts.
- Warehouses are ~163 km from the production hub on average, showing geographic spread.
- Most warehouses were established around 2009; missing wh_est_year values need imputation.
- Variables like govt checks, breakdowns, and temperature regulation show low frequencies but may still impact efficiency.

Data Preprocessing

1. Drop irrelevant Columns

I. Ware_house_ID
II. WH_Manager_ID

Mentioned columns, containing only distinct identification values, offered no analytical value for the project's objective of understanding product quantity. Consequently, these columns are removed.

2. Missing Values

As we can observe-

```
Feature name | No. of null values =  % of null values

workers_num                | 990  = 3.96  %
wh_est_year                | 11881 = 47.52 %
approved_wh_govt_certificate | 908  = 3.63  %
```

TABLE 4 – MISSING VALUES

Rishabh Gupta

Variable '***wh_est_year'***, is having almost half of the entries missing. Replacing almost half the data with a single value will significantly reduce the variance of this feature and could distort its relationship with the target variable. This could lead to biased model training. **Hence, we will drop the variable from our analysis.**

Rest two of the columns, we will impute the values.

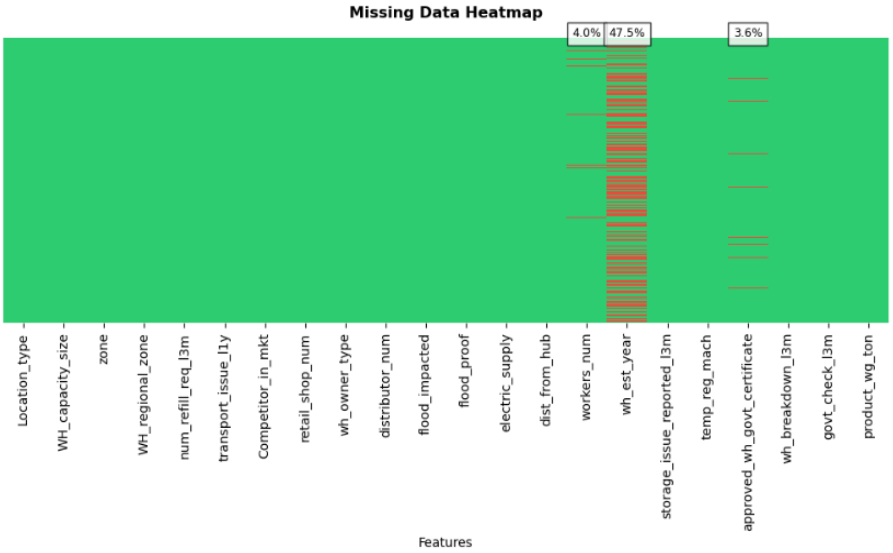


FIGURE 1 - MISSING VALUES MAP

Here, **workers_num :** numerical column, thus we will impute missing values with median.

approved_wh_govt_certificate: categorical column, thus we will impute missing values with mode.

Finally, we got,

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
wh_owner_type	0
distributor_num	0
flood_impacted	0
flood_proof	0
electric_supply	0
dist_from_hub	0
workers_num	0
storage_issue_reported_l3m	0
temp_reg_mach	0
approved_wh_govt_certificate	0
wh_breakdown_l3m	0
govt_check_l3m	0
product_wg_ton	0
dtype: int64	

TABLE 5 – MISSING VALUES TREATMENT

3. **Duplicate Values** – There are no duplicate values in dataset.

Rishabh Gupta

4. Outliers Treatment

Outliers were present in some of the columns, lets plot boxplots to locate outliers and do the necessary treatment.

We have plotted 17 distplot and 17 boxplots for our eligible variables. Please refer code.
Out of them, we can observe that below two variables are unique and contains only 1 value.
Thus, it will not contribute to our analysis.

i. **flood_impacted**
ii. **flood_proof**

Outliers were detected and removed.

Exploratory Data analysis

Let's analyze Numerical variables-

Numerical variables summary:

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt
count	25000.000000	25000.000000	25000.000000
mean	4.089040	0.773680	3.104200
std	2.606612	1.199449	1.141663
min	0.000000	0.000000	0.000000
25%	2.000000	0.000000	2.000000
50%	4.000000	0.000000	3.000000
75%	6.000000	1.000000	4.000000
max	8.000000	5.000000	12.000000

	retail_shop_num	distributor_num	dist_from_hub	workers_num
count	25000.000000	25000.000000	25000.000000	24010.000000
mean	4985.711560	42.418120	163.537320	28.944398
std	1052.825252	16.064329	62.718609	7.872534
min	1821.000000	15.000000	55.000000	10.000000
25%	4313.000000	29.000000	109.000000	24.000000
50%	4859.000000	42.000000	164.000000	28.000000
75%	5500.000000	56.000000	218.000000	33.000000
max	11008.000000	70.000000	271.000000	98.000000

	wh_breakdown_l3m	govt_check_l3m	product_wg_ton
count	25000.000000	25000.000000	25000.000000
mean	3.482040	18.812280	22102.632920
std	1.690335	8.632382	11607.755077
min	0.000000	1.000000	2065.000000
25%	2.000000	11.000000	12059.000000
50%	3.000000	21.000000	22101.000000
75%	5.000000	26.000000	30103.000000
max	6.000000	32.000000	55151.000000

Rishabh Gupta

TABLE 6 – NUMERICAL COLUMN SUMMARY

- Missing values in workers_num (990 entries) need attention via imputation or exclusion.
- Skewness varies: some features are slightly or highly skewed, while others like product_wg_ton are fairly symmetric.
- High variability in product_wg_ton and retail_shop_num points to differing demand and retail reach across warehouses.
- Wide ranges in key variables like retail_shop_num reflect diverse operational scales. Low transport issues for most warehouses, but variability in refills and retail presence suggests demand-supply imbalance.

Let's analyze Categorical variables –

- **Rural Bias:** The dataset is dominated by rural warehouses, which may skew results toward rural-specific trends.
- **Capacity Distribution**: Most warehouses are "Large" or "Mid" capacity; "Small" ones are underrepresented.
- **East Zone Limitation:** Very few entries from the East zone hinder region-specific insights there.
- **Detailed Zoning:** WH_regional_zone provides finer regional granularity, useful for analyzing demand patterns.

- **Ownership & Certification Balance:** A fairly even mix of ownership types and certifications suggests diverse yet balanced warehouse profiles.

```
Unique values in Location_type:
Location_type
Rural    22957
Urban     2043
Name: count, dtype: int64

Unique values in WH_capacity_size:
WH_capacity_size
Large    10169
Mid      10020
Small     4811
Name: count, dtype: int64

Unique values in zone:
zone
North    10278
West      7931
South     6362
East       429
Name: count, dtype: int64

Unique values in WH_regional_zone:
WH_regional_zone
Zone 6    8339
Zone 5    4587
Zone 4    4176
Zone 2    2963
Zone 3    2881
Zone 1    2054
Name: count, dtype: int64

Unique values in wh_owner_type:
wh_owner_type
Company Owned    13578
Rented           11422
Name: count, dtype: int64

Unique values in approved_wh_govt_certificate:
approved_wh_govt_certificate
C     5501
B+    4917
B     4812
A     4671
A+    4191
Name: count, dtype: int64
```

Table 7 – Categorical Column summary

Rishabh Gupta

Univariate Analysis

➢ **Continuous variables** – Let's plot Histogram and boxplot for each continues variables and analyses

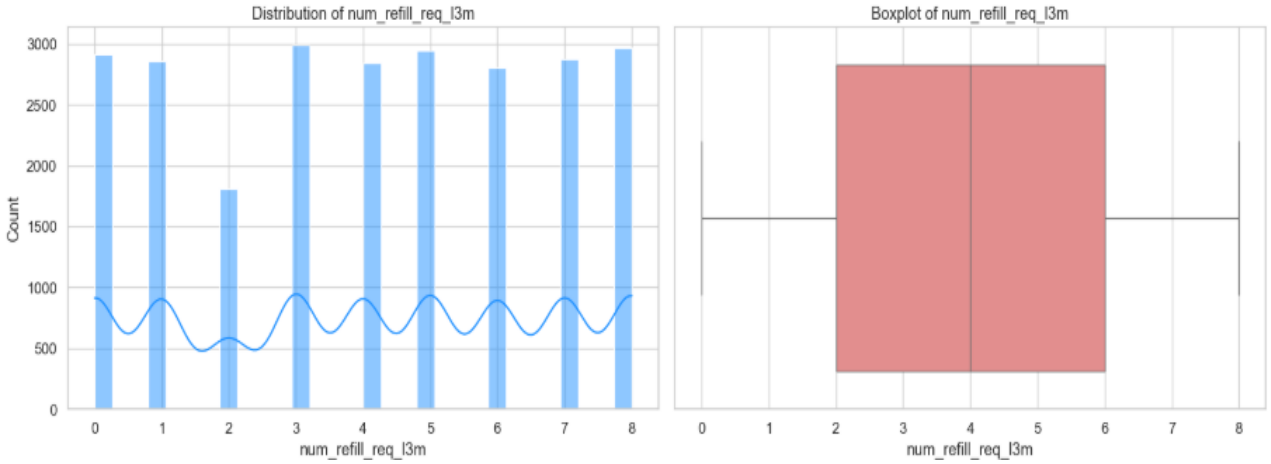


FIGURE 2 – UNIVARIATE ANALYSIS FOR NUMBER OF TIMES REFILLING DONE

- Refill Patterns: Refill requests (num_refill_req_l3m) show a discrete, peaked distribution with counts

mostly between 2 and 6.

- Consistent Range: Median is 4, and the data spans 0 to 8 without outliers, indicating stable refill behavior.
- Operational Insight: The discrete, regular pattern hints at standardized inventory practices across warehouses.

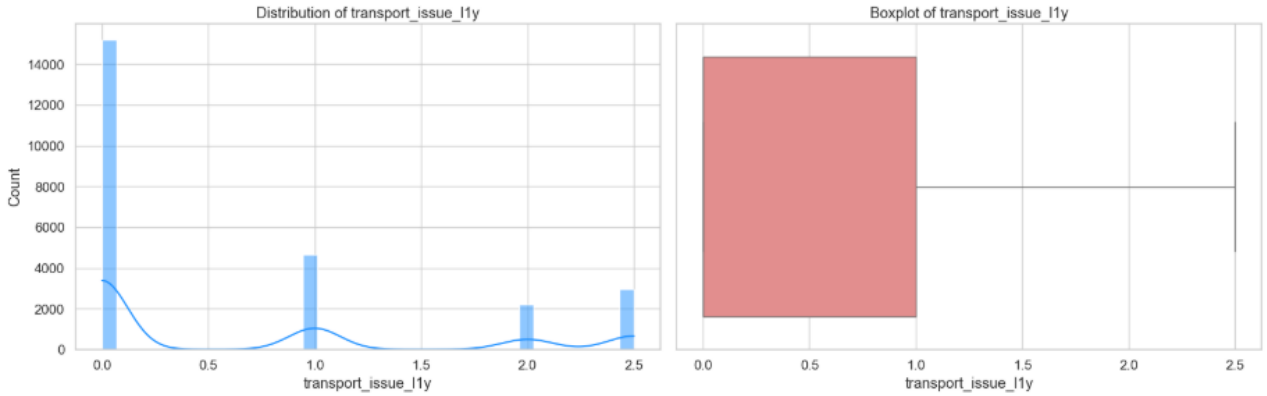


FIGURE 3 - UNIVARIATE ANALYSIS FOR TRANSPORT ISSUE

- Right-Skewed Pattern: Most warehouses reported zero transport issues, with a sharp histogram peak at 0.
- Rare Incidents: Fewer entries at 1–3 issues suggest transport problems are uncommon but not absent.

Rishabh Gupta

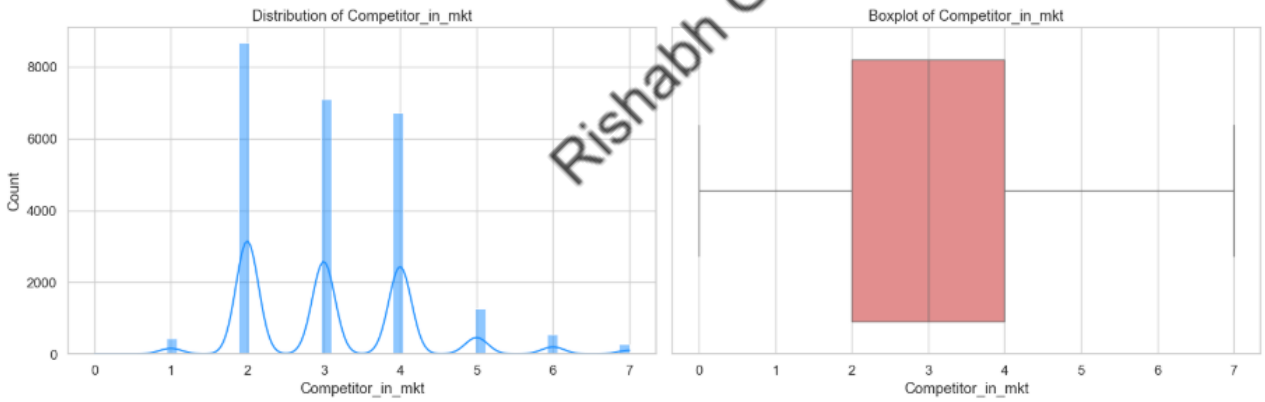


FIGURE 4 - UNIVARIATE ANALYSIS FOR COMPETITION IN MKT

- Discrete Distribution: Competitor_in_mkt shows clear peaks at whole numbers, reflecting distinct competitor counts per market.
- Moderate Concentration: Median is around 3, with most values between 2 and 4, indicating limited variation.
- Segmentation Insight: Varying competitor levels hint at potential market segmentation affecting warehouse demand.

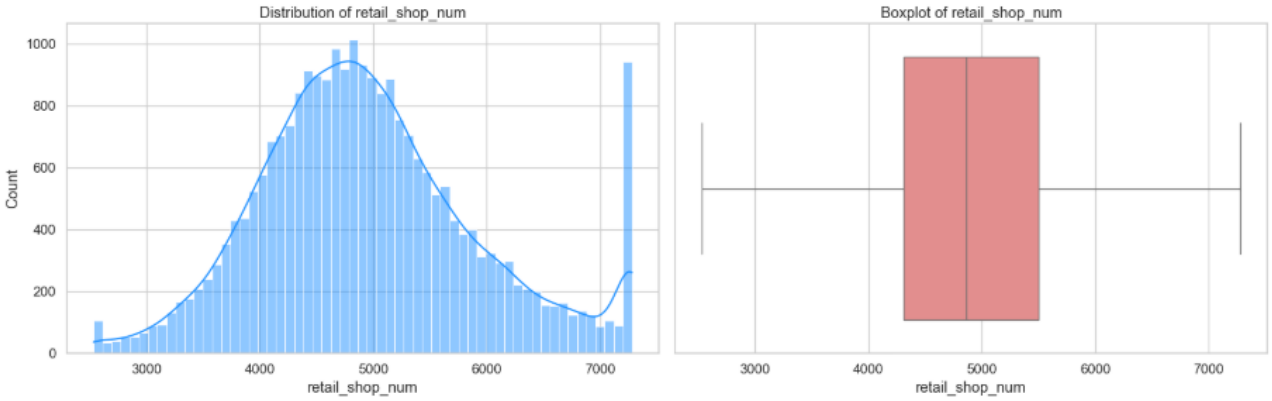


FIGURE 5 - UNIVARIATE ANALYSIS FOR RETAIL SHOPS QUANTITY

- Wide Distribution: retail_shop_num shows a broad, roughly normal spread with a peak near 5000 shops.
- High Variability: A large interquartile range and outliers indicate major differences in retail reach across warehouses.
- Demand Driver: This variation likely impacts warehouse demand, making it a key variable for supply planning.

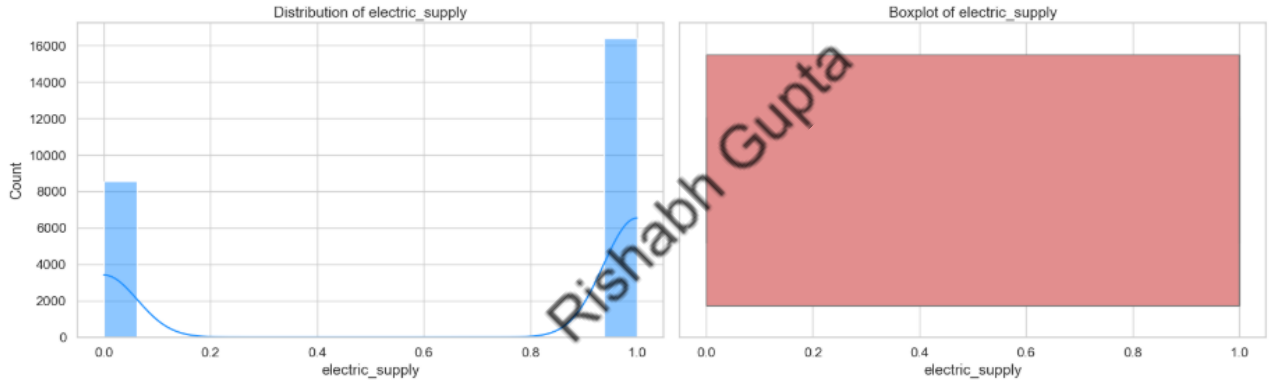


FIGURE 6 - UNIVARIATE ANALYSIS FOR ELECTRIC SUPPLY

- Binary Nature: electric_supply shows two clear values (0 and 1), with most warehouses having electricity.
- Skewed Distribution: A dominant peak at 1 indicates widespread availability of electric backup.
- Operational Relevance: This feature may significantly impact warehouse functionality and storage capabilities.

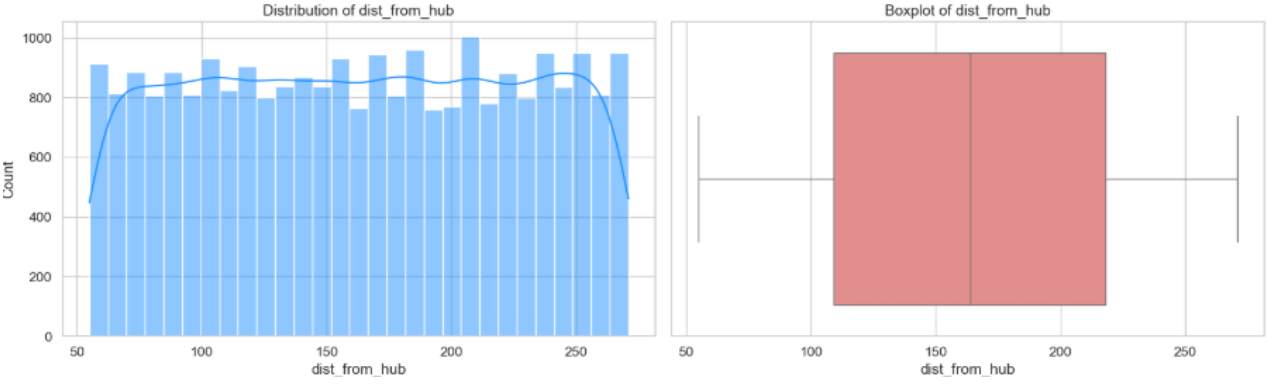


FIGURE 7 - UNIVARIATE ANALYSIS FOR DISTANCE FROM HUB

- Even Spread: dist_from_hub shows a fairly uniform distribution, indicating warehouses are geographically dispersed.
- Moderate Variability: The interquartile range is balanced, with no major outliers in distance.
- Logistics Impact: This spread may lead to variation in transportation costs and delivery times

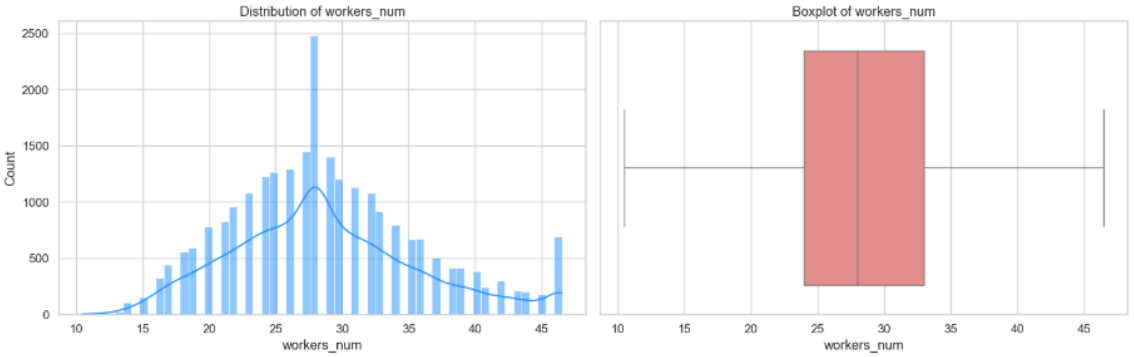


FIGURE 8 - UNIVARIATE ANALYSIS FOR WORKERS NUMEBRS

- Bimodal & Skewed: workers_num shows two peaks (around 28–30 and 45), with a slight right skew, indicating varied staffing patterns.
- Moderate Variation: Most warehouses have 24–33 workers, but some outliers suggest larger teams.
- Capacity Indicator: Worker count likely reflects operational scale and different warehouse models.

Rishabh Gupta

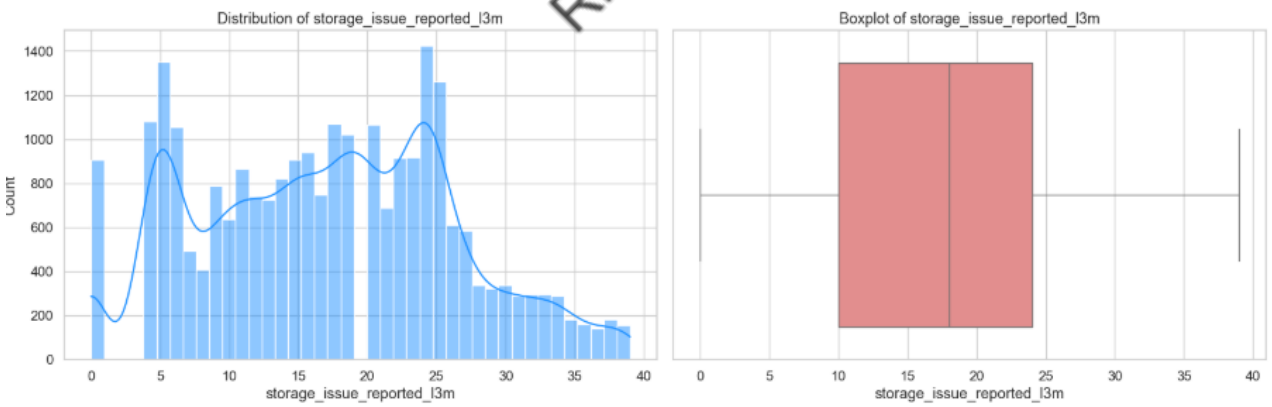


FIGURE 9 - UNIVARIATE ANALYSIS FOR STORAGE ISSUE REPORTED

- **Multi-Modal Peaks:** storage_issue_reported_l3m shows several common levels of reported issues, with peaks near 0, 7-8, and 25-26.
- **Right-Skewed:** The median is around 13, with a longer upper tail indicating some warehouses report many more issues.
- **Operational Differences:** These patterns suggest varying storage management efficiencies across warehouses.

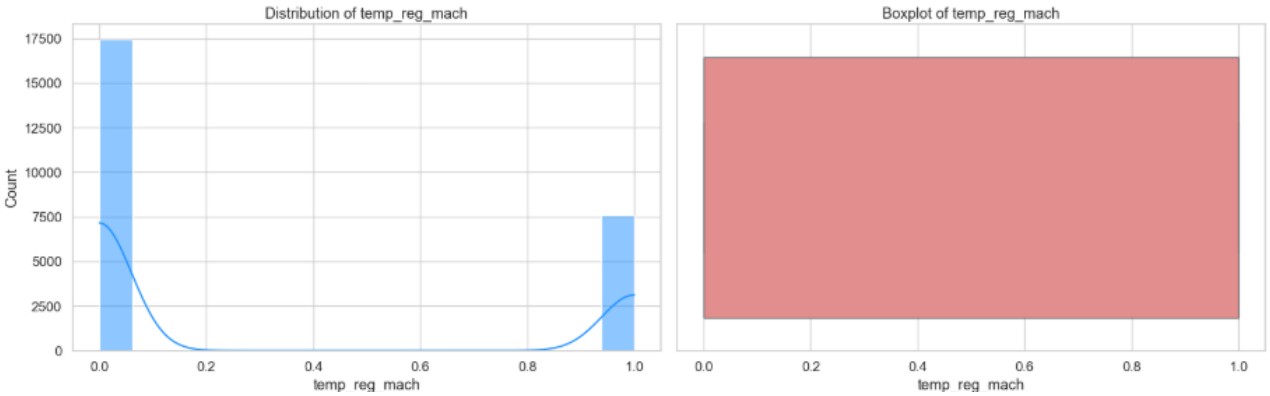
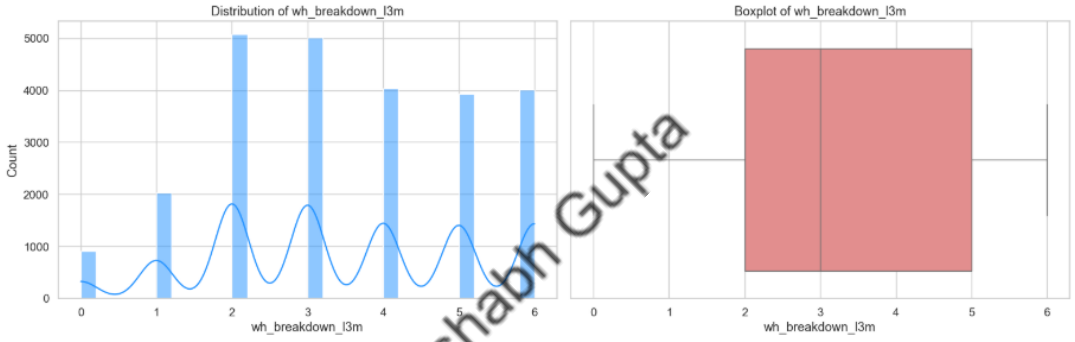


FIGURE 10 - UNIVARIATE ANALYSIS FOR TEMPERATURE

- Binary Variable: temp_reg_mach shows two clear values (0 and 1), indicating presence or absence of temperature regulation.
- Mostly Absent: A higher peak at 0 suggests most warehouses lack this machinery.
- Quality Impact: Lack of temperature control may affect product quality and shelf life.



Rishabh Gupta

FIGURE 11 - UNIVARIATE ANALYSIS FOR WAREHOUSE BREAKDOWN

- Discrete Counts: wh_breakdown_l3m shows whole-number breakdowns with distinct histogram peaks.
- Right Skewed: Median is around 3, with some warehouses experiencing up to 5 or 6 breakdowns.
- Operational Impact: Frequent breakdowns may affect supply reliability and inventory management at certain warehouses.

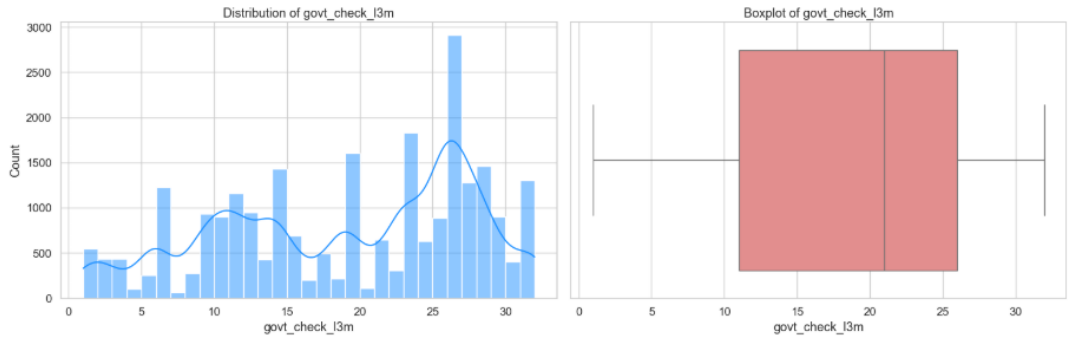


FIGURE 12 - UNIVARIATE ANALYSIS FOR GOVERNMENT CHECKS

- Multi-Modal & Right-Skewed: govt_check_l3m shows multiple peaks with more warehouses having higher numbers of government checks.
- Median & Spread: Median is around 21, with most checks falling between 11 and 26.

- Regulatory Variation: Differences in government inspections likely reflect varying regulatory scrutiny or compliance needs across warehouses.

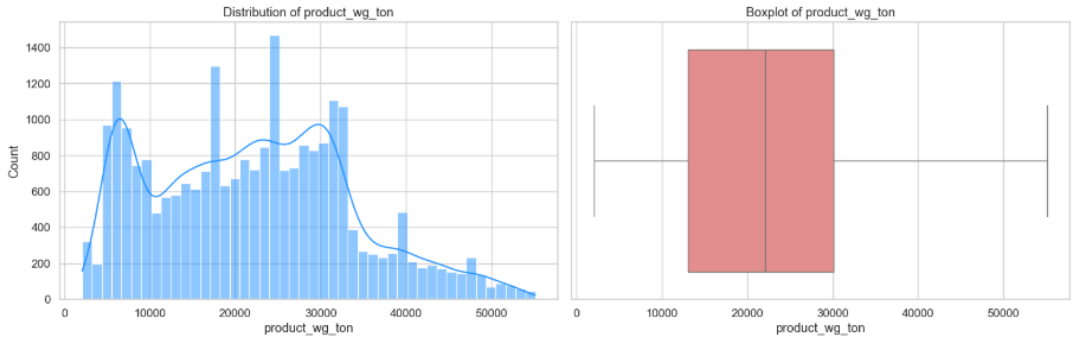


FIGURE 13 - UNIVARIATE ANALYSIS OR PRODUCT WEIGH

- Bimodal & Right-Skewed: product_wg_ton has two peaks around 7,000-8,000 and 25,000-30,000 tons, with a right-skewed distribution.
- Median & Range: Median shipment weight is about 22,000 tons, with most values between 13,000 and 30,000 tons.
- Wide Spread & Outliers: A long upper whisker suggests some unusually large shipments, highlighting demand-supply mismatches across warehouses.

Rishabh Gupta

➢ **Categorical variables –** Let's plot countplot for each categorical variables and analyses.

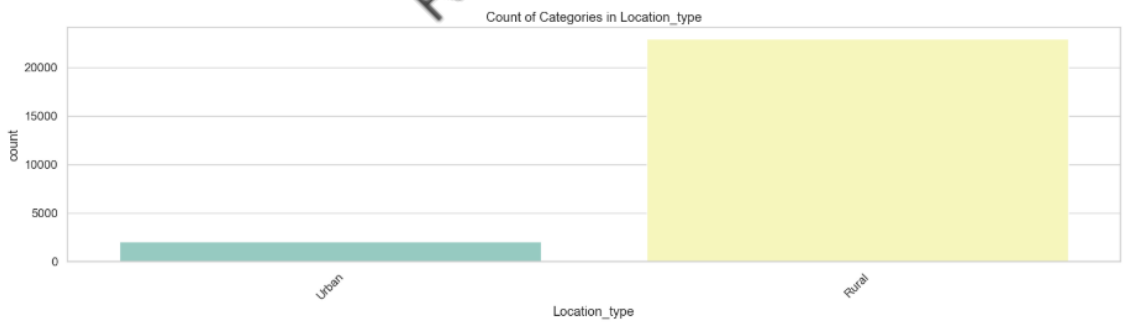


FIGURE 14 - UNIVARIATE ANALYSIS FOR LOCATION TYPE

- **Dominant Rural Presence:** The dataset contains over 22,000 warehouses in rural areas versus just over 2,000 in urban locations.
- **Imbalanced Representation:** This significant imbalance may bias models toward rural patterns, underrepresenting urban dynamics.
- **Stratification Needed:** To ensure fair model training and testing, stratified sampling by Location_type is recommended.

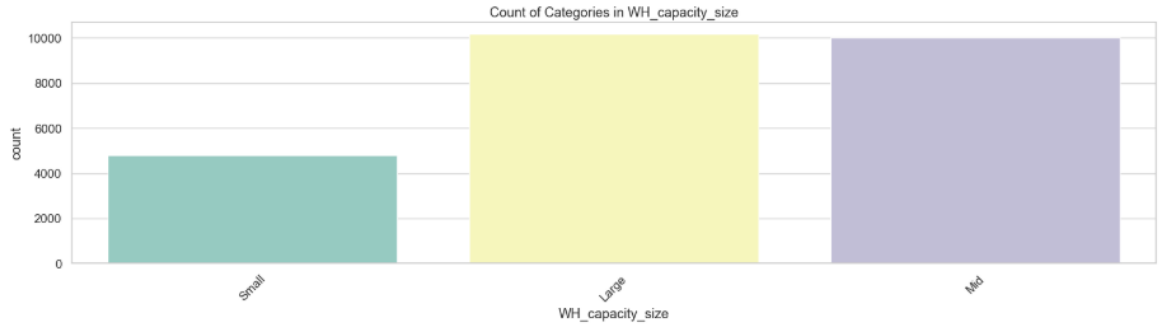
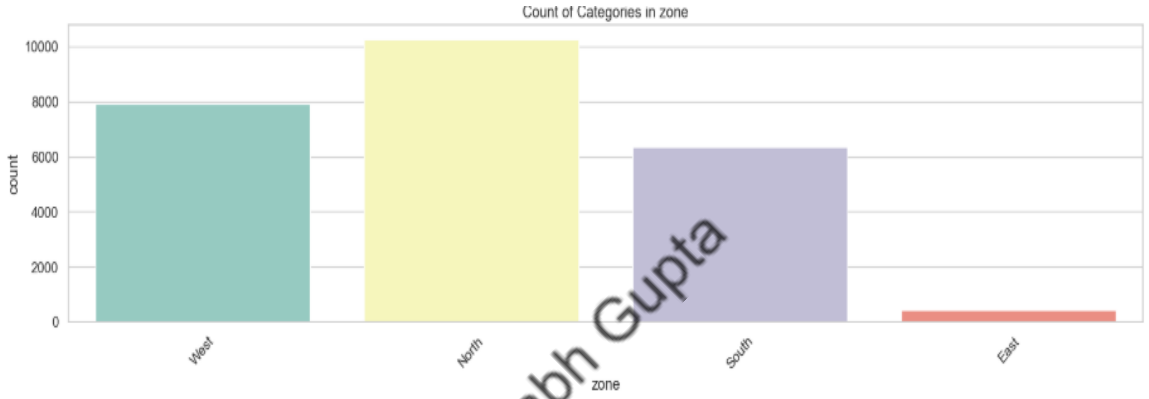


FIGURE 15 - UNIVARIATE ANALYSIS FOR WHEREHOUSE CAPACITY

- Balanced Large and Mid: "Large" and "Mid" capacity warehouses have similarly high counts in the dataset.
- Fewer Small Warehouses: The "Small" capacity category is noticeably less represented.
- Size-Based Analysis Potential: This distribution supports meaningful comparisons across warehouse sizes for operational or supply pattern insights.
-



Rishabh Gupta

FIGURE 16 - UNIVARIATE ANALYSIS FOR ZONES

- Dominant North and West Zones: Most warehouses are located in the "North" and "West" zones.
- Moderate South Zone: The "South" zone has fewer warehouses than the North and West but a moderate presence.
- Underrepresented East Zone: The "East" zone has very few warehouses, limiting reliable analysis for that region.

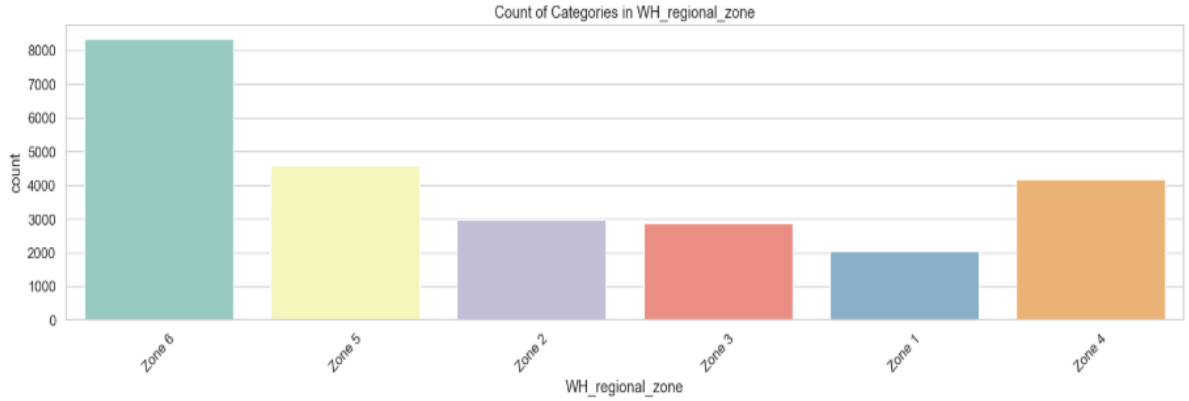


FIGURE 17 - UNIVARIATE ANALYSIS FOR WEARHOUSE REGIONAL ZONES

- Zone 6 Dominance: "Zone 6" has the highest number of warehouses among regional zones.
- Varied Representation: Other zones ("Zone 5", "Zone 2", "Zone 3", "Zone 1", "Zone 4") show much lower and varied counts.
- Detailed Insights: This granular breakdown enables targeted analysis of demand and supply patterns for specific regional strategies.

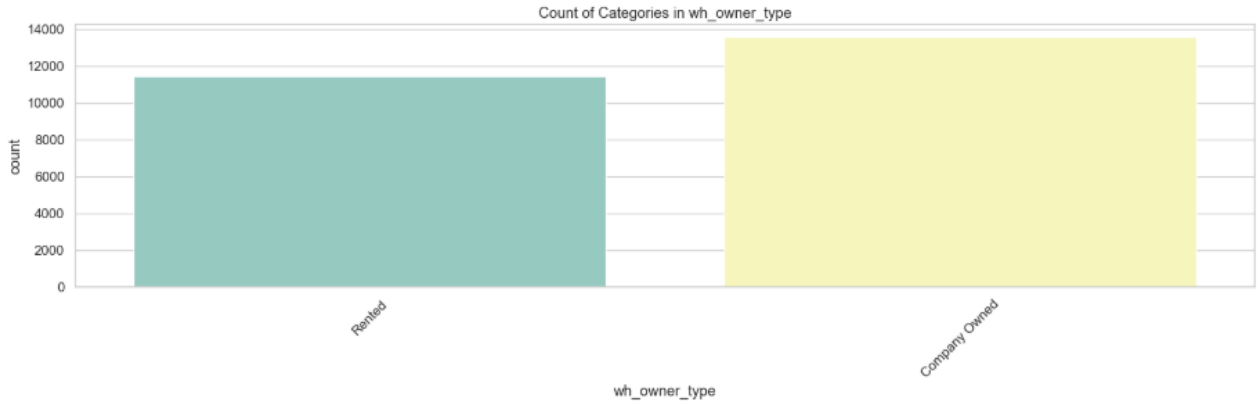


FIGURE 18 - UNIVARIATE ANALYSIS FOR OWNER TYPE

- Majority Company-Owned: "Company Owned" warehouses slightly outnumber "Rented" ones.
- Reasonable Balance: Both ownership types are well represented in the dataset.
- Segmented Analysis Potential: Ownership may impact operations, so analyzing them separately could reveal important supply chain differences.

Rishabh Gupta

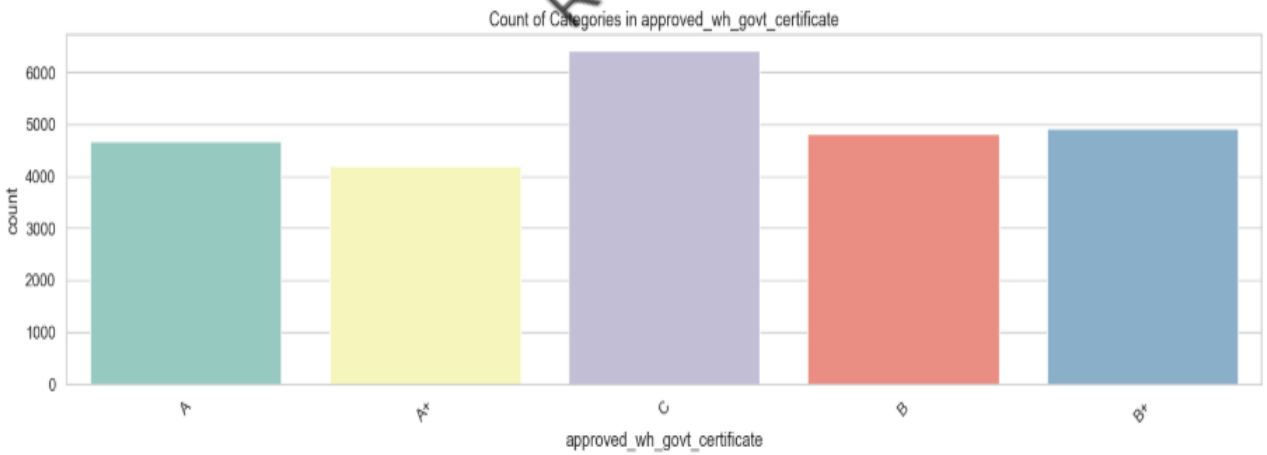
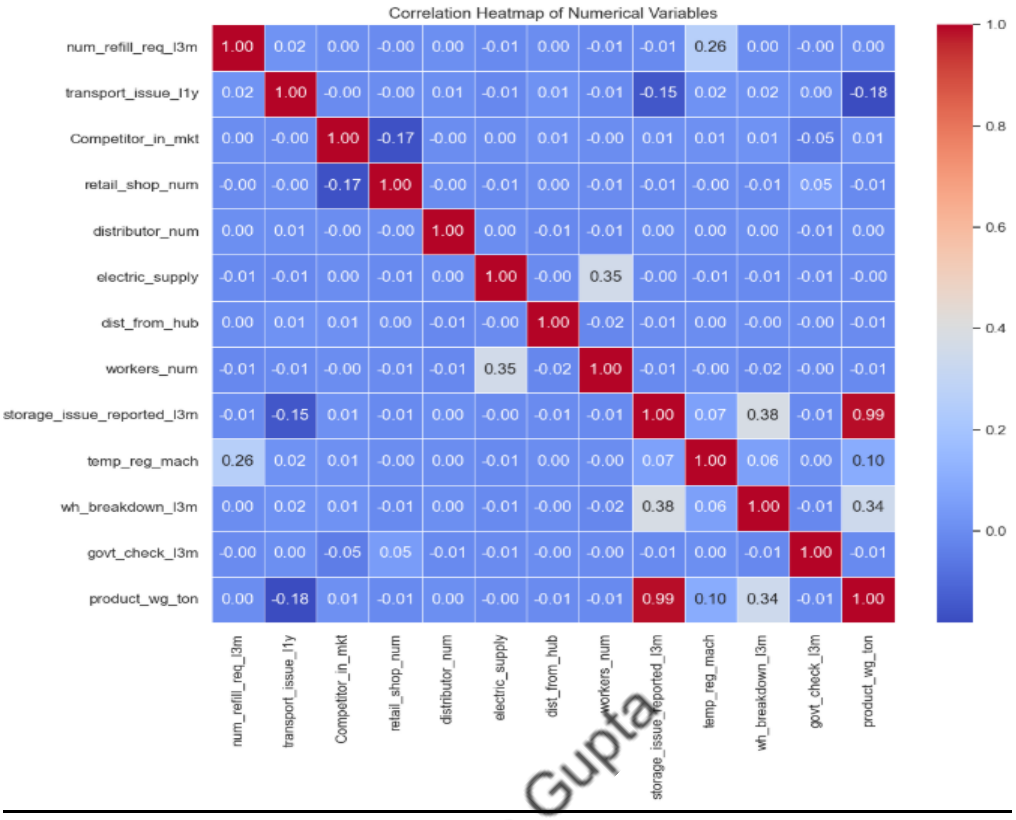


FIGURE 19 - UNIVARIATE ANALYSIS FOR GOVT CERTIFICATE

- Even Distribution: Warehouse counts across government certificate categories ("A", "A+", "C", "B", "B+") are fairly balanced.
- Slightly Higher 'C' Category: The "C" category has a marginally larger share of warehouses.
- Operational Correlation Potential: Exploring links between certificate levels and factors like efficiency, storage issues, or shipment volume could provide valuable insights.

Bivariate Analysis

➢ Correlation Heatmap (Numerical Variables)



Rishabh Gupta

FIGURE 20 - CORRELATION HEATMAP(NUMERICAL VARIABLES)

- workers_num and storage_issue_reported_l3m have a very strong positive correlation (0.99), while temp_reg_mach moderately correlates (0.26) with num_refill_req_l3m.
- product_wg_ton shows weak correlations with workers_num (0.10), temp_reg_mach (0.34), and transport_issue_l1y (-0.18), with most other features weakly correlated.
- High correlation between workers_num and storage_issue_reported_l3m may cause multicollinearity, and weak overall linear correlations suggest non-linear models could better predict product_wg_ton.

➢ Scatter Plots (Numerical vs Target)

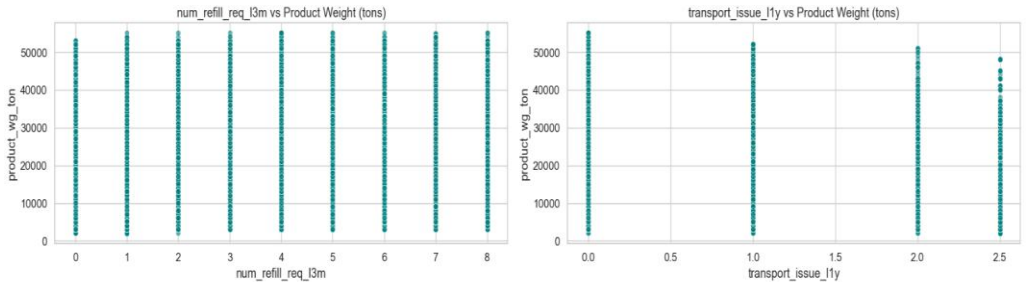


Figure 21 - Scatter plot for Refills done and Transport issues

- The left scatter plot shows distinct vertical bands for num_refill_req_l3m, confirming its discrete

nature, with no clear linear link to product_wg_ton.

- The right scatter plot clusters mostly at transport_issue_l1y = 0, showing no obvious linear trend but allowing high product weights even without transport issues.
- Both plots suggest no strong linear correlation with product_wg_ton, indicating possible non-linear effects or influences from other variables, with wide variance in product weight across values.

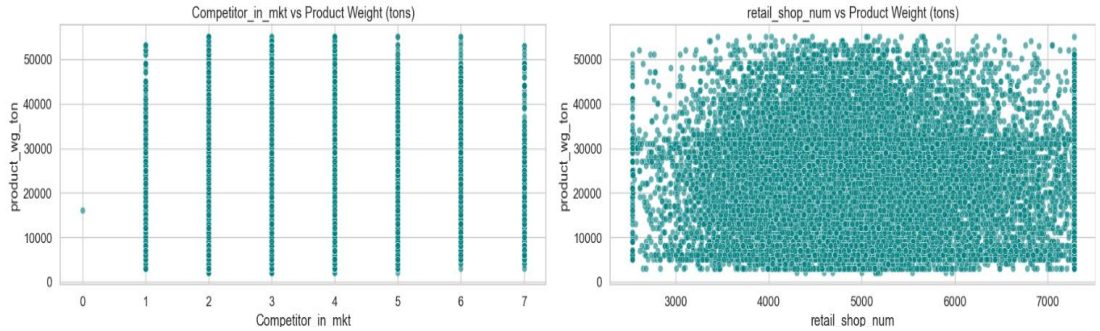


FIGURE 22 - SCATTERPLOT FOR RETAIL SHOPS AND COMPETITORS

- The left scatter plot shows discrete vertical bands for competitor counts, with no clear linear link to product_wg_ton and a wide spread of shipment weights for each count.
- The right scatter plot of retail_shop_num versus product_wg_ton shows no distinct linear pattern but suggests higher retail counts may allow for larger maximum shipment weights.
- Significant variance in product_wg_ton exists across all competitor and retail shop counts, indicating other factors also play important roles.

Rishabh Gupta

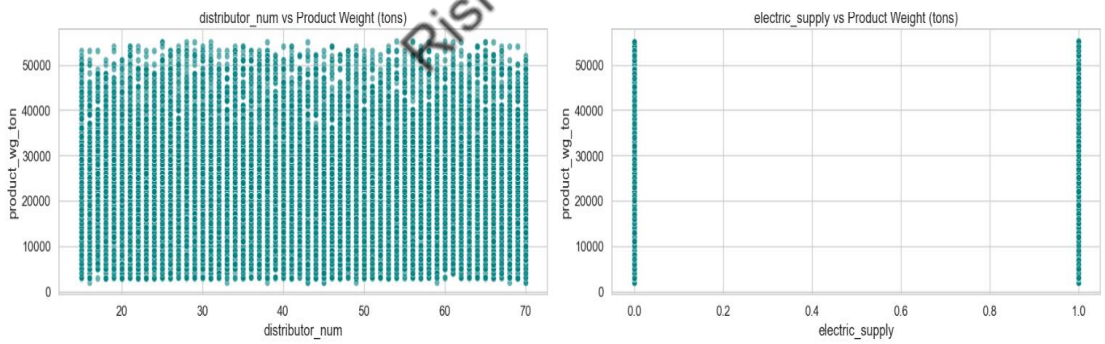


FIGURE 23 - SCATTER PLOT FOR DISTRIBUTERS AND ELECTRIC SUPPLY

- The left scatter plot shows no clear trend between distributor_num and product_wg_ton, indicating distributor count alone isn't a strong predictor of shipment weight.
- The right scatter plot confirms electric_supply is binary, with product_wg_ton widely spread across both supply categories.
- Although electric supply doesn't directly affect shipment weight, it may influence other operational factors that impact product_wg_ton, suggesting a need for multivariate analysis.

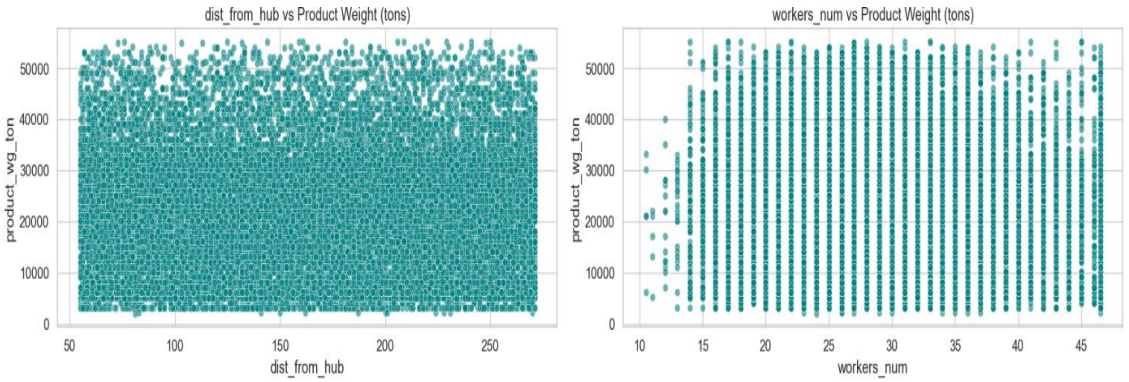
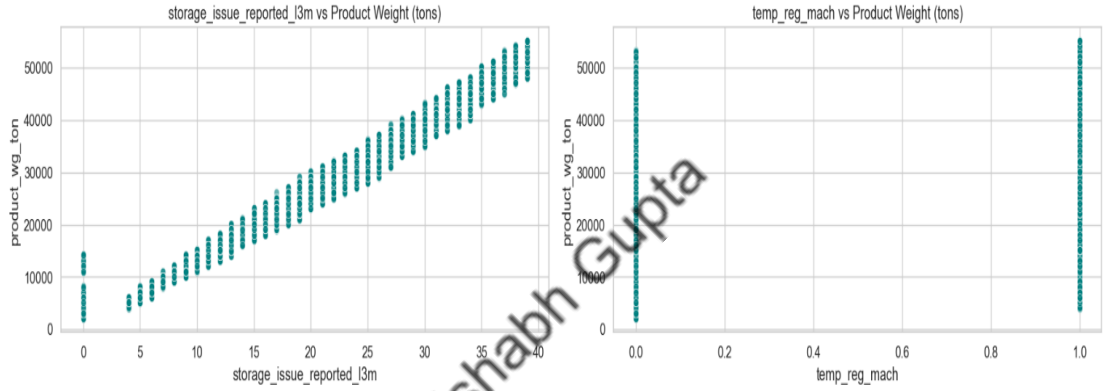


FIGURE 24 - SCATTER PLOT FOR DISTRIBUTER FROM HUB AND WORKERS NUMBERS

- The left scatter plot shows no clear linear relationship between dist_from_hub and product_wg_ton, with wide variation in shipment weight across all distances.
- The right scatter plot indicates a weak positive trend between workers_num and product_wg_ton, with higher worker counts generally linked to higher shipment weights.
- Despite this trend, there is significant variation in product_wg_ton for each worker count, suggesting other factors also influence shipment weight.



Rishabh Gupta

FIGURE 25 - SCATTER PLOT FOR STORAGE ISSUE REPORTED AND TEMPERATURES

- The left scatter plot shows a positive correlation between storage_issue_reported_l3m and product_wg_ton, with higher storage issues linked to larger shipment weights.
- The right scatter plot confirms temp_reg_mach is binary, and higher product weights tend to occur more with temperature regulation, though large shipments also exist without it.
- These patterns suggest that warehouses handling bigger volumes face more storage challenges, and temperature control may be important for maintaining product quality in larger shipments.

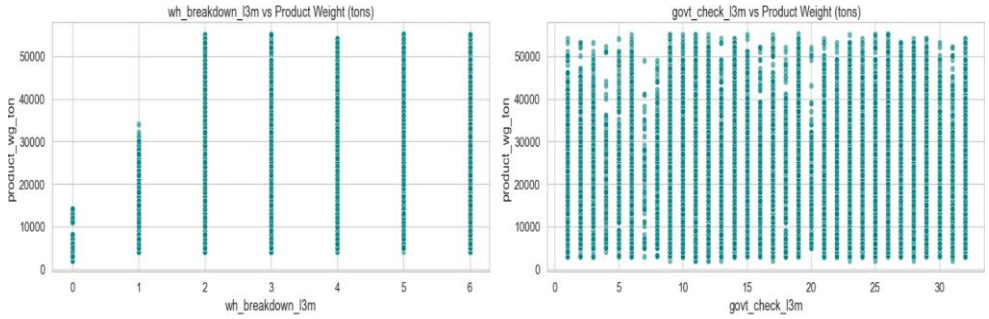


FIGURE 26 - SCATTERPLOT FOR WAREHOUSE BREAKDOWNS AND GOVT CHECKS

- The left scatter plot shows discrete warehouse breakdowns (wh_breakdown_l3m) with no clear linear trend relating breakdowns to product_wg_ton, though breakdowns may cause varied shipment sizes during recovery.

- The right scatter plot displays discrete government checks (govt_check_l3m) also without a clear linear correlation to shipment weight.
- Both breakdowns and government checks may indirectly affect supply chain efficiency and shipment patterns, but wide variance in product_wg_ton suggests other factors play a stronger role.

Boxplots (Categorical vs Target)

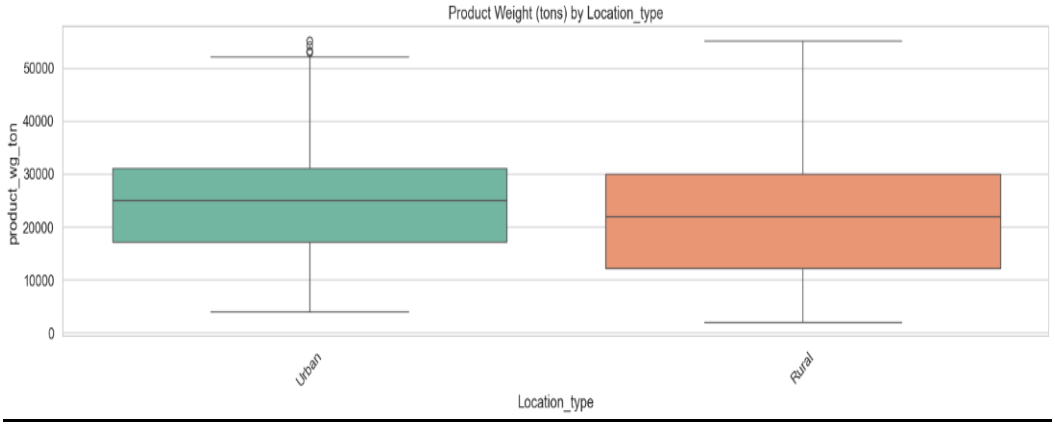


FIGURE 27 - BOXPLOT - LOCATION VS TARGET

- Median product weight (product_wg_ton) is similar for both Urban and Rural locations.
- Rural warehouses show greater variability and higher maximum shipment weights compared to Urban ones.
- Urban locations have more high-value outliers, indicating differing supply strategies between the two areas.

Rishabh Gupta

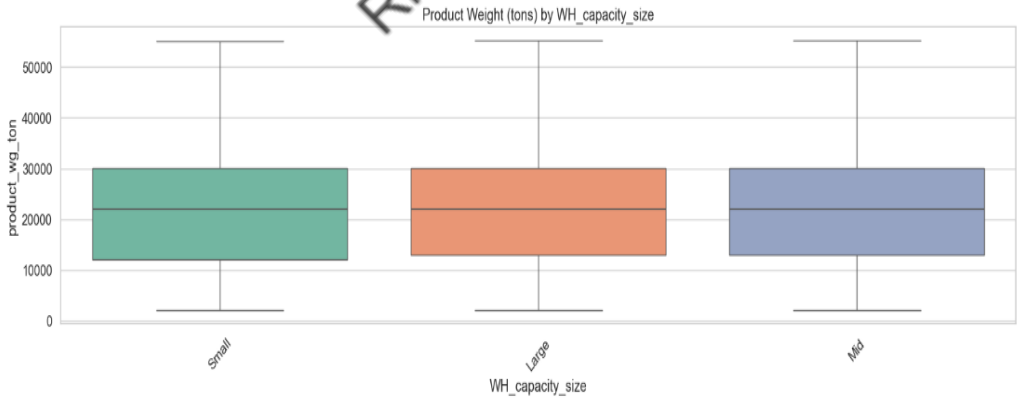


FIGURE 28 - BOXPLOT - WAREHOUSE CAPACITY VS TARGET

- Median product weight (product_wg_ton) is fairly consistent across Small, Mid, and Large warehouses.
- Large warehouses show less variability in shipment sizes, suggesting more standardized operations.
- Small and Mid warehouses have wider shipment weight spreads, indicating more variability despite similar maximum shipment sizes.

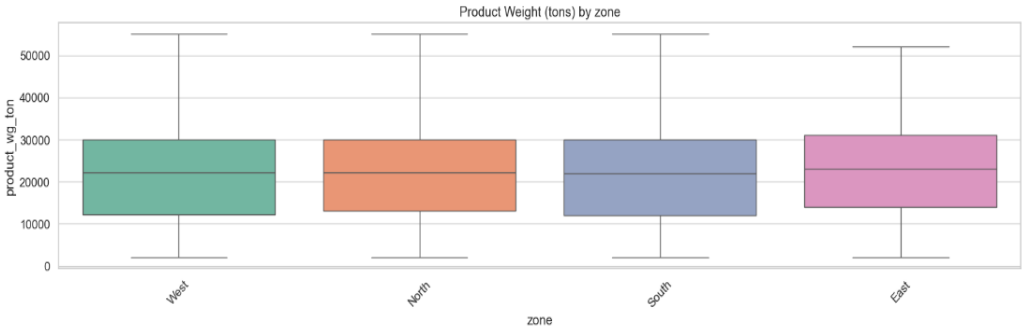


FIGURE 29 - BOXPLOT - CAPACITY VS TARGET

- Median product weight (product_wg_ton) is fairly consistent across all four zones (West, North, South, East).
- The East zone shows the highest variability in shipment weights, while West and South have moderate, similar spreads.
- Large shipments occur in all zones, but regional factors may affect the consistency of shipment sizes rather than the median weight itself.

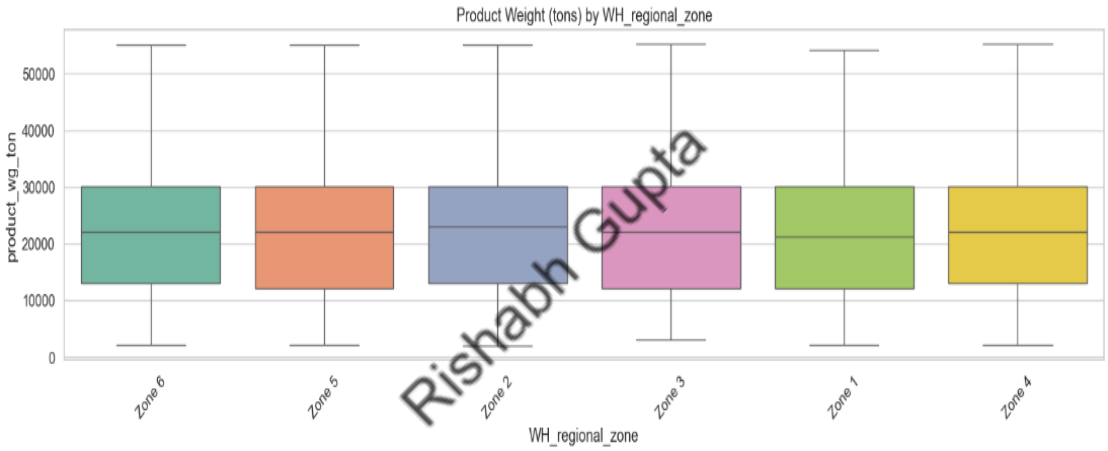


FIGURE 30 - BOXPLOT - WH REGIONAL ZONE VS TARGET

- Median product_wg_ton is consistent across all six regional zones.
- Zone 5 shows the greatest variability in shipment weights, while Zones 1 and 4 have more consistent shipment sizes.
- Large shipments occur in all zones, and variability differences suggest targeted inventory strategies may be needed, especially for Zone 5.

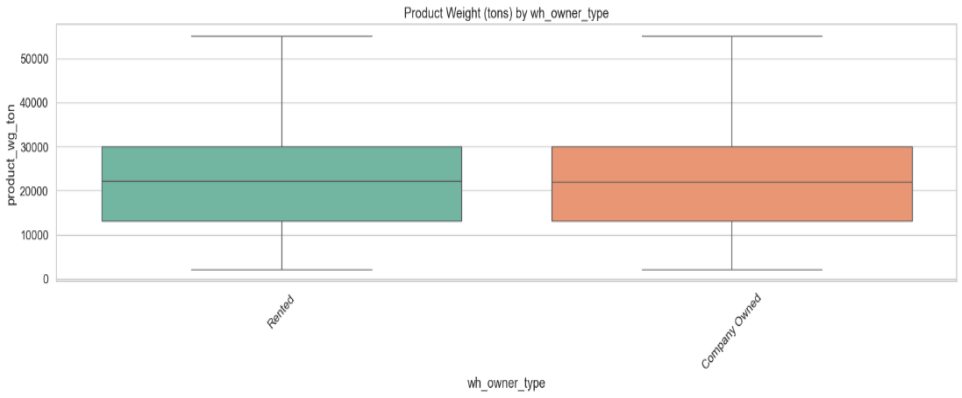


FIGURE 31 - BOXPLOT - OWNER TYPE VS TARGET

- Median product_wg_ton is similar for both "Rented" and "Company Owned" warehouses.
- "Rented" warehouses show slightly more variability in shipment weights than company-owned ones.
- Large shipments occur in both types, indicating ownership isn't a key factor in median shipment weight but may influence operational variability.

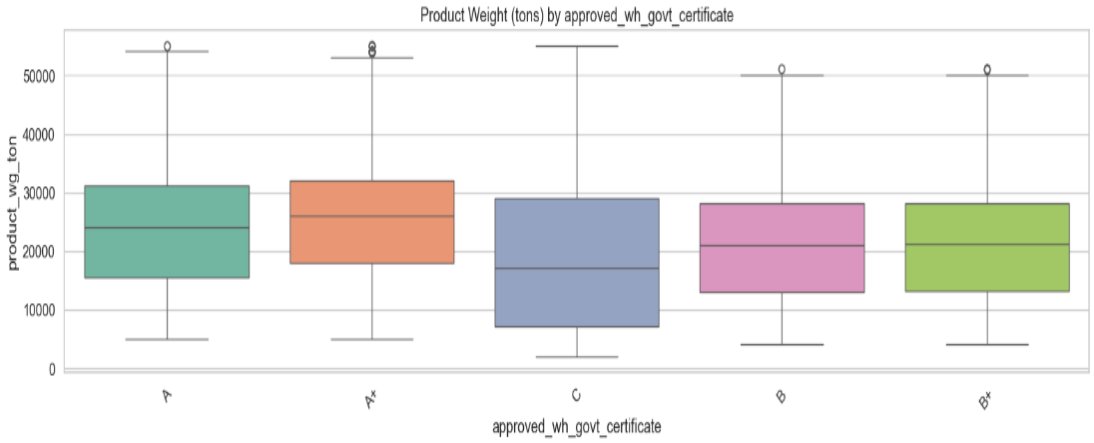


FIGURE 32 - BOXPLOT - APPROVED GOVT CERT VS TARGET

- Median product_wg_ton varies by government certificate, with 'C' lowest and 'A+' highest.
- 'C' certified warehouses show the least variability, while 'A+' warehouses have the widest spread in shipment weights.
- Large shipments occur across all certificates, suggesting higher-tier certifications may relate to larger-scale operations, needing further study.

Rishabh Gupta

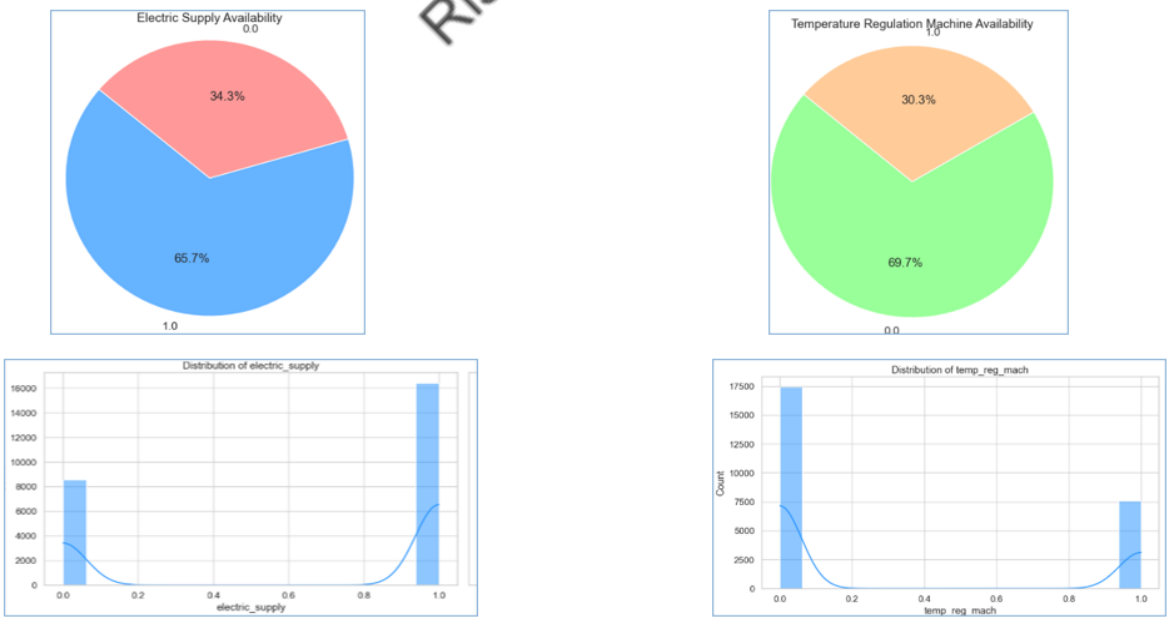


FIGURE 33 - PIECHART AND BAR CHART COMPARISON

- Both **temp_reg_mach downtime** and **electricity shortages** drastically reduce average product weight, with outages cutting output by **~75-80%**.

Multivariate Analysis

Count Heatmap

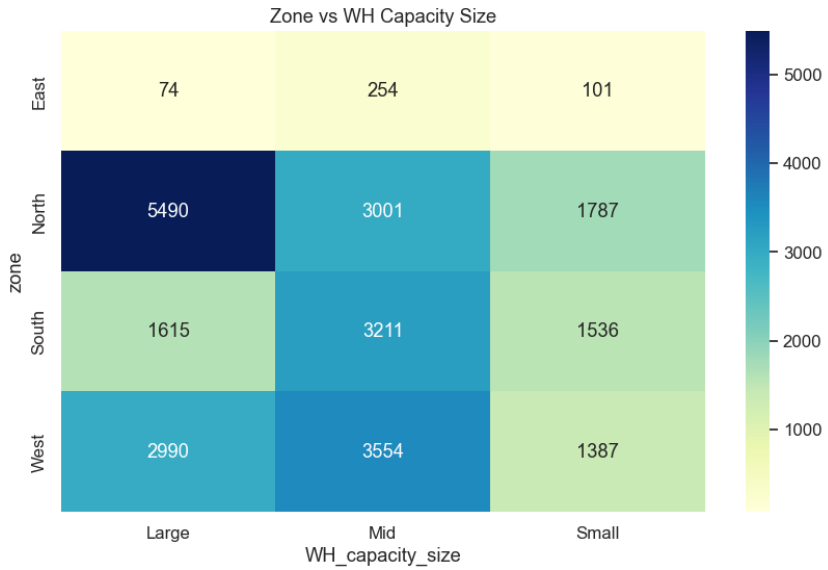


FIGURE 34 - COUNT HEATMAP FOR ZONE VS CAPACITY SIZE

- North Zone leads in warehouse count, especially for large capacities, while East Zone has the fewest warehouses across all sizes.
- Capacity distribution varies by zone: North dominates in large warehouses; West and South lead in mid-sized; East shows a balanced but low count.
- Warehouse size concentration suggests strategic placement by region, so regional analyses must consider these capacity and zone disparities, particularly East's underrepresentation.

Rishabh Gupta

Violin Plot - Combines boxplot and KDE to show the distribution of product_wg_ton across categories like zone

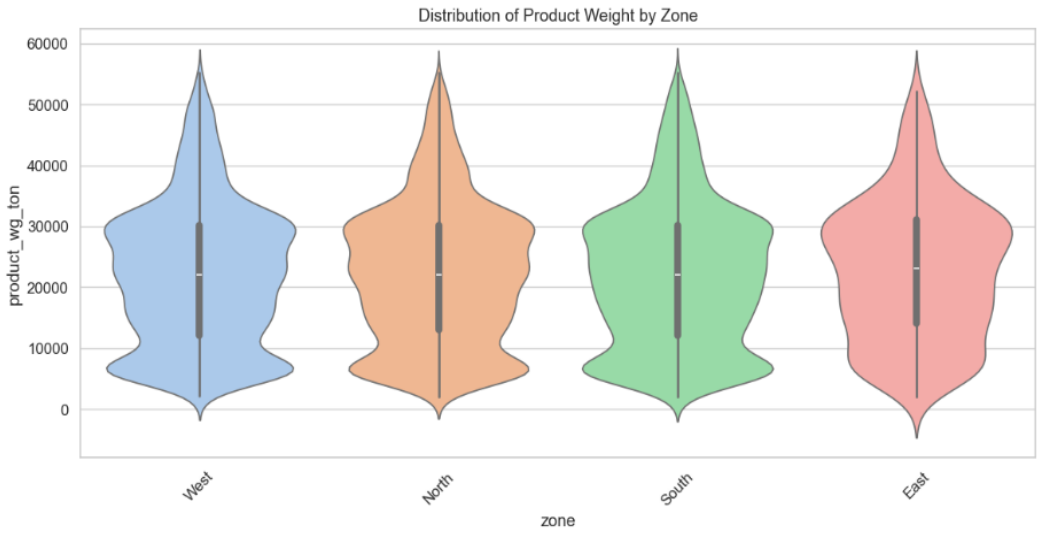


FIGURE 35 - VIOLIN PLOT

- Median shipment weights are similar across West, North, South, and East zones, showing consistent average shipment sizes.

- North and West zones display greater variability in shipment weights, while South shows more stability; East has highly consistent shipments but with limited data.

FacetGrid Bar Plot - Visualizes product_wg_ton means across categories using subplots category.

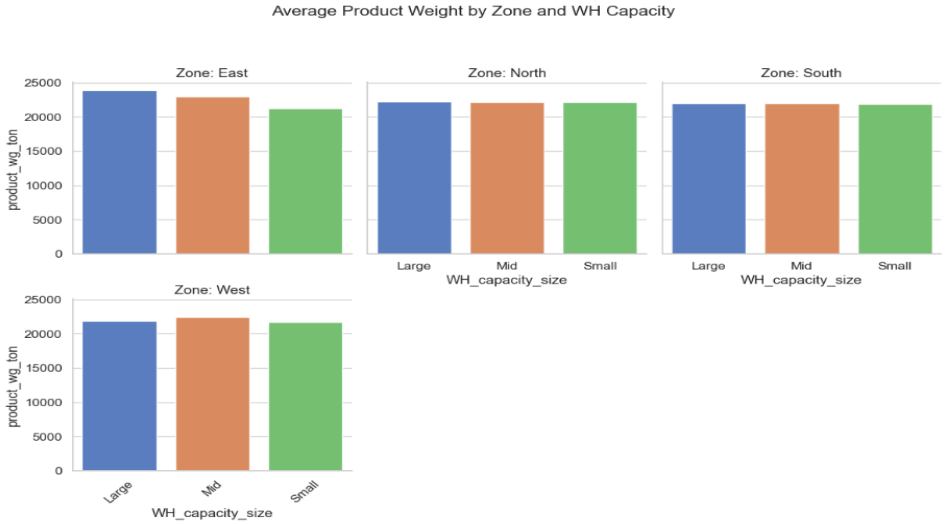


FIGURE 36 - FACETGRID BAR PLOT

- Within each zone (East, North, South, West), average product weight (product_wg_ton) is similar across warehouse capacities (Large, Mid, Small), showing capacity size isn't a key driver within zones.
- Minor variation exists between zones, with the East zone having a slightly higher average product weight, possibly due to zone-specific demand or logistics factors.
- The East zone's higher average, despite fewer warehouses, suggests zone-level factors may influence shipment weight more than capacity size, highlighting the need for zone-focused supply chain strategies.

Rishabh Gupta

Pair Plot (with Hue) - Visualizes relationships among multiple numeric variables with class differentiation using colour (hue).

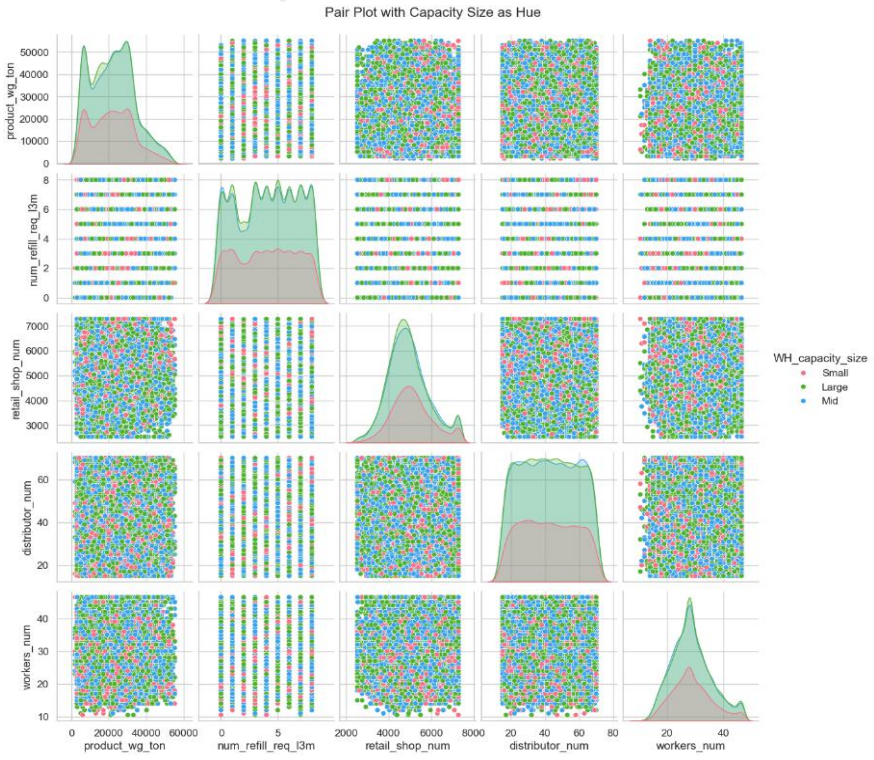


FIGURE 37 - PAIRPLOT

- Larger warehouses (“Large,” “Mid”) handle a wider and higher range of product weights, serve more retail shops, and have more workers; “Small” warehouses handle lower weights and fewer shops/workers.
- Refill requests are discrete across capacities, with slightly higher counts in “Large” and “Mid” warehouses; number of distributors shows no clear link to capacity.
- Weak positive trends exist between product weight, retail shops, workers, and refill requests, highlighting the need for multivariate analysis to understand shipment weight drivers.
-

Bubble Plot - Scatter plot where point size encodes a third numerical variable and colour indicates category

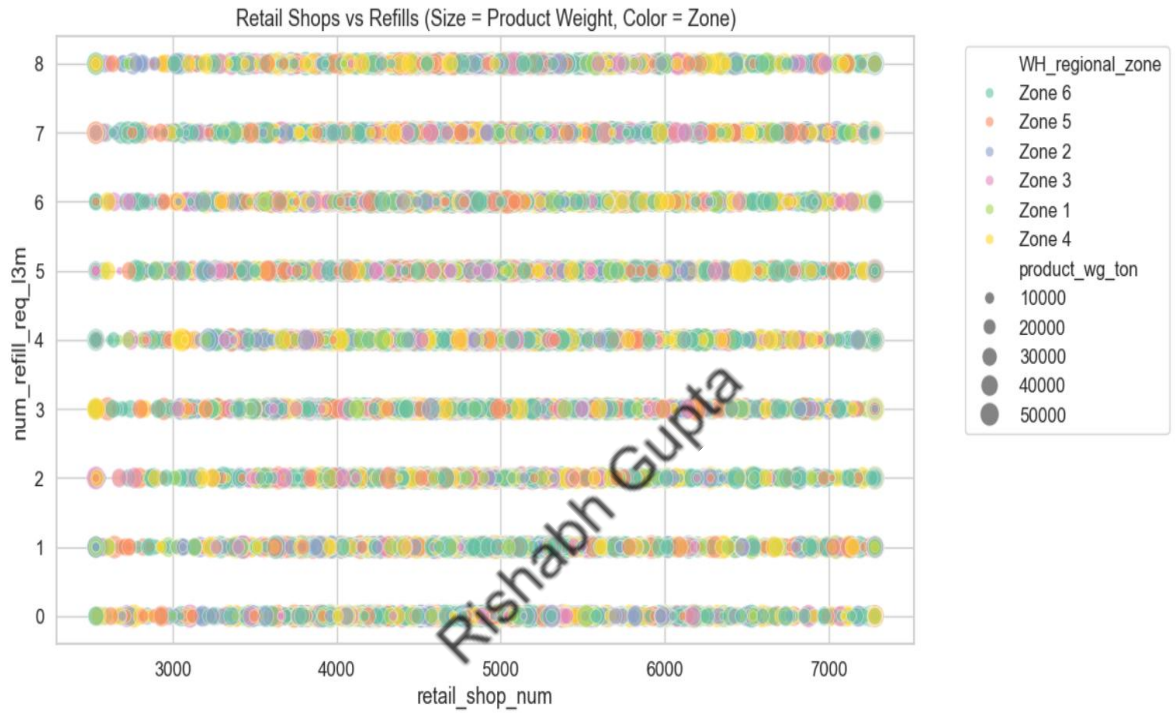


FIGURE 38 - BUBBLE PLOT

- The number of refill requests (num_refill_req_l3m) is a discrete variable ranging from 0 to 8, while the number of retail shops served varies widely from about 2500 to 7500 across all refill levels.
- There is no clear linear relationship between product weight (product_wg_ton) and either refill requests or retail shop count; large shipments occur at various levels of both.
- All regional zones are represented across the data, with no strong bivariate correlation between refill requests and retail shops, though subtle regional patterns may exist.

Parallel Coordinates Plot - Visually track how different features contribute to the product_wg_ton, highlighting Location_type.

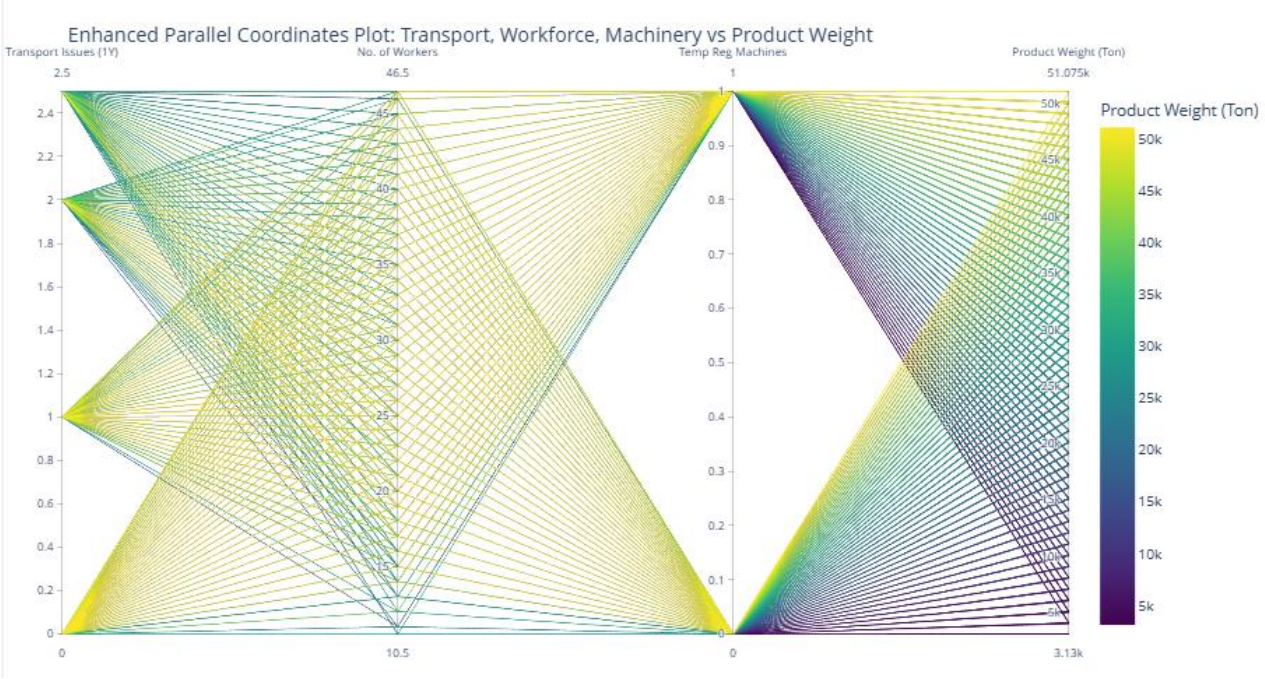


FIGURE 39 - PARALLEL COORDINATES PLOT

- Warehouses with temperature regulation generally ship heavier product weights, especially when combined with a larger workforce.
- Workforce size shows a positive association with shipment volume, while transport issues do not have a clear direct impact on product weight.
- Temperature control and workforce are key influencers of shipment weight, indicating complex interactions among these factors.

Rishabh Gupta

3D Scatter Plot – Lastly, let explore the relationship between dist_from_hub, electric_supply, and product_wg_ton, segmented by approved_wh_govt_certificate.

It is very complex to analyse but we will have a look.

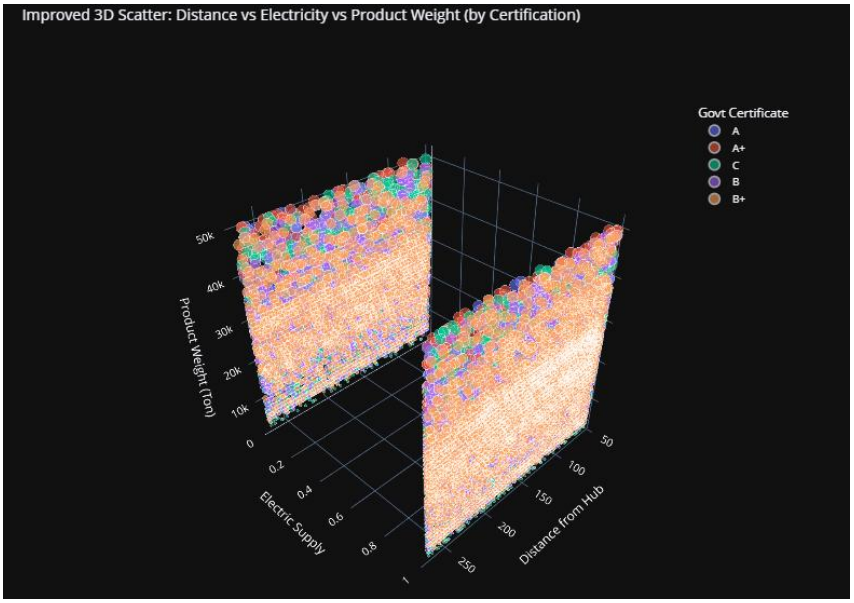


FIGURE 40 - 3D SCATTER PLOT

- Electric Supply is a binary variable, clearly split between absence (0) and presence (1).
- Product weights vary widely, showing significant shipment size diversity across warehouses.
- Warehouses are spread over various distances from the hub, with more clustered near shorter distances.
- All government certificate levels appear across different combinations of electric supply, distance, and product weight.
- No clear linear relationship is seen in 3D, though subtle, complex patterns or regional tendencies by certificate might exist.

Business insights from EDA

- Demand & Supply Drivers: Warehouses with more retail shops indicate stronger demand; optimize forecasting and refill batches. Transport issues are uneven; model distance and transport problems for logistics resilience.
- Warehouse Characteristics: Majority are Large or Mid capacity, balanced ownership; investigate if ownership affects efficiency. Certifications are well-distributed; use as proxies for quality.
- Location Insights: Rural warehouses dominate; North zone leads in numbers, East is underrepresented—consider tailored strategies and expansion opportunities.
- Operational Factors: Workforce varies by capacity; storage issues exist in some warehouses; few have temperature control—assess cost-benefit for quality improvements. Interaction between temperature control and workforce links to higher shipment weights.
- Target Variable: Product weight is bimodal and right-skewed, signaling demand-supply mismatch; use non-linear models (e.g., XGBoost) and optimize shipment sizes.
- Clustering Insights: Identify clusters like high-demand hubs, efficient operators, remote locations, and problematic warehouses; tailor strategies accordingly. Use k-means or hierarchical clustering with key variables for segmentation.
- Additional Business Insights: Evaluate temperature control expansion, workforce planning aligned with shipment volume, and East zone underrepresentation. Analyze transport issues and enforce compliance to minimize disruptions.
- Regional Demand Variations: North zone shows highest median shipments; East zone is a potential growth market needing focused analysis.
- Data Imbalance: Skewed rural-urban and zone distribution risks biased models; apply stratified sampling, oversampling, weighted models, and targeted data collection.
- Strategic Recommendations: Develop segment-specific models, benchmark operations across clusters, and prioritize infrastructure investments like temperature control in high-volume warehouses.

Rishabh Gupta

Model Building

➢ Feature Selection

After data preprocessing and EDA, we have below mentioned feature columns with us on which we will proceed with model building.

#	Column	Non-Null Count	Dtype
0	Location_type	25000 non-null	object
1	WH_capacity_size	25000 non-null	object
2	zone	25000 non-null	object
3	WH_regional_zone	25000 non-null	object
4	num_refill_req_l3m	25000 non-null	float64
5	transport_issue_l1y	25000 non-null	float64
6	Competitor_in_mkt	25000 non-null	float64
7	retail_shop_num	25000 non-null	float64
8	wh_owner_type	25000 non-null	object
9	distributor_num	25000 non-null	float64
10	electric_supply	25000 non-null	float64
11	dist_from_hub	25000 non-null	float64
12	workers_num	25000 non-null	float64
13	storage_issue_reported_l3m	25000 non-null	float64
14	temp_reg_mach	25000 non-null	float64
15	approved_wh_govt_certificate	25000 non-null	object
16	wh_breakdown_l3m	25000 non-null	float64
17	govt_check_l3m	25000 non-null	float64
18	product_wg_ton	25000 non-null	float64

dtypes: float64(13), object(6)

TABLE 8 - FEATURE SELECTIONS AFTER DATA PREPROCESSING

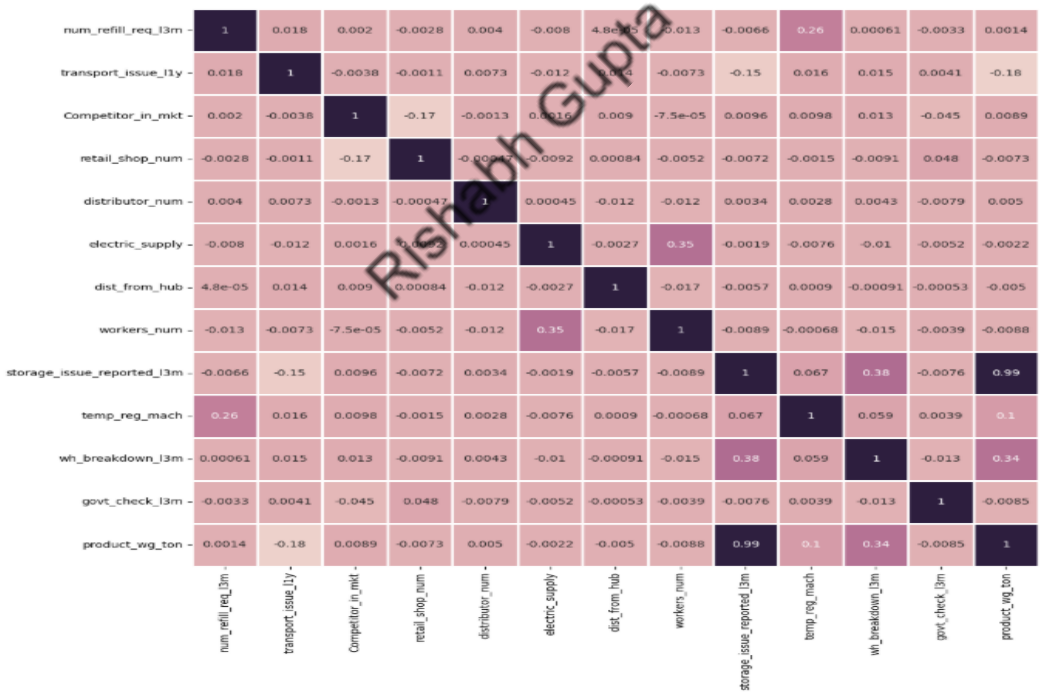


FIGURE 41 - CORRELATION MAP FOR FEATURES

Key Observations:

- Strong positive correlations exist between storage issues and product weight (0.99), temperature regulation and product weight (0.9), and warehouse breakdowns and storage issues (0.67).
- Moderate correlations include electric supply with workforce (0.35), temperature regulation with storage issues (0.38), and storage issues with government checks (0.38); weak negative correlations appear between transport issues and product weight (-0.18), and competitor presence with retail

shops (-0.17).

- Many other variable pairs show weak or near-zero correlations, indicating little linear relationship; the high correlation between storage issues and product weight suggests potential multicollinearity.

➢ Encoding

We applied ***One-Hot Encoding*** to below categorical columns because of their nominal nature-

- Location_type
- zone
- WH_regional_zone
- WH_owner_type

And we applied Label encoding to below features due to their ordinal nature-

- WH_capacity
- approved_wh_govt_certificate
- So, after encoding we get –

```
Number of data points :  25000
Number of features/columns :  25
 ------------------------------------------------------------------------------
Features :  ['WH_capacity_size' 'num_refill_req_l3m' 'transport_issue_l1y'
 'Competitor_in_mkt' 'retail_shop_num' 'distributor_num' 'electric_supply'
 'dist_from_hub' 'workers_num' 'storage_issue_reported_l3m'
 'temp_reg_mach' 'approved_wh_govt_certificate' 'wh_breakdown_l3m'
 'govt_check_l3m' 'product_wg_ton' 'Location_type_Urban' 'zone_North'
 'zone_South' 'zone_West' 'WH_regional_zone_Zone 2'
 'WH_regional_zone_Zone 3' 'WH_regional_zone_Zone 4'
 'WH_regional_zone_Zone 5' 'WH_regional_zone_Zone 6'
 'wh_owner_type_Rented']
```

Rishabh Gupta

TABLE 9 – DATA AFTER ENCODING

➢ Standardisation after Train Test Split

We use standardization before training for most machine learning models — especially those that are sensitive to feature scale, such as: Linear, SVM etc.
Please refer code.

Now we will build models and compare actual vs predicted values.

Data Modelling

1. Linear Regression Model

To find the line (or hyperplane in higher dimensions) that minimizes the **sum of squared residuals (RSS)** between the predicted values and the actual values in the training data.

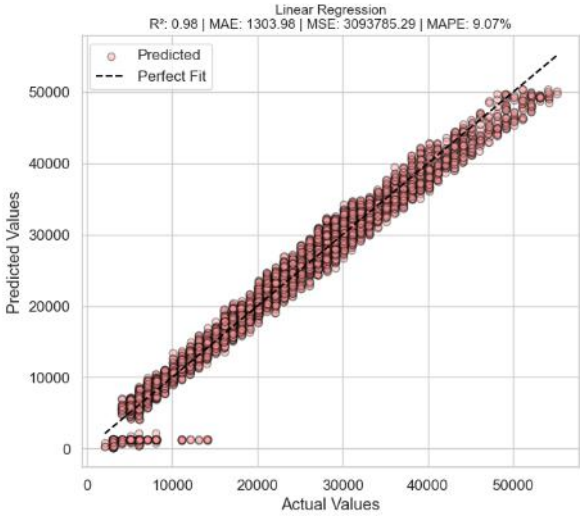


FIGURE 42 - LINEAR REGRESSION MODEL PLOT

- The scatter plot shows a strong positive linear relationship between predicted and actual values, with points closely clustered around the diagonal, indicating accurate model predictions.
- The model's error metrics include a Mean Absolute Error (MAE) of about 1304 and a Mean Squared Error (MSE) of roughly 3.09 million, reflecting average prediction errors and emphasizing larger deviations.
- With an R-squared of approximately 0.977, the model explains 97.7% of the variance, indicating a very good fit, though some scattered points at lower values suggest occasional larger prediction errors.

Rishabh Gupta

2. Ridge Regression (L2 Regularization)

To minimize the sum of squared residuals **plus a penalty term** proportional to the **square of the magnitude of the coefficients**. This penalty discourages large coefficient values.

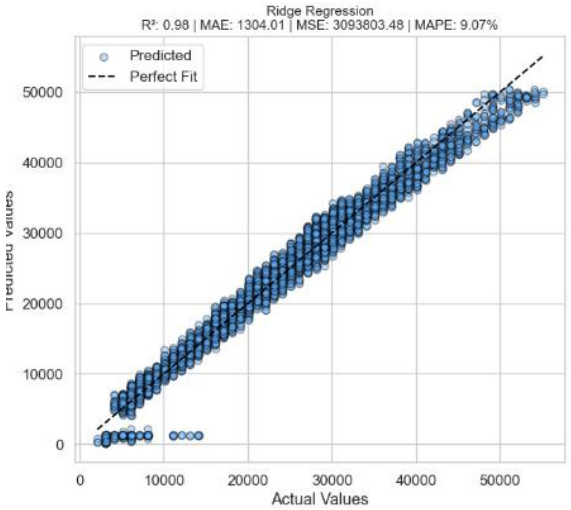


FIGURE 43 - RIDGE REGRESSION PLOT

- The scatter plot for Ridge Regression shows a strong positive linear relationship between predicted and actual values, closely matching the Linear Regression plot and indicating a good fit.
- Error metrics are nearly identical to Linear Regression, with MAE around 1304.01 and MSE approximately 3.094 million, reflecting similar average and squared prediction errors.
- The R-squared value of about 0.9769 confirms that Ridge Regression explains a comparable amount of variance, suggesting L2 regularization had minimal impact on performance, likely due to limited multicollinearity in the data.

3. Decision Tree Model

This uses a tree-like structure to classify data based on feature splits.

- The scatter plot shows a strong positive correlation between predicted and actual values, though with slightly more spread at lower predictions compared to linear models, indicating more variability there.
- Decision Tree Regressor significantly improves error metrics, with MAE dropping to ~869.62 and MSE to ~1.74 million, reflecting smaller average and squared errors than linear models.
- The R-squared increases to ~0.987, showing the tree's ability to capture complex, non-linear relationships, with top splits prioritizing storage_issue_reported_l3m to progressively reduce error and refine predictions.

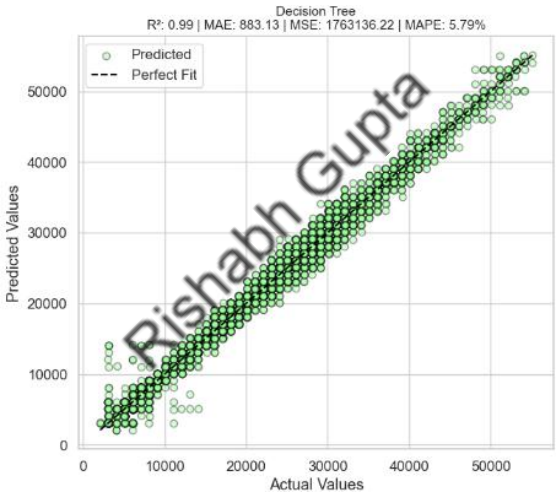


FIGURE 44 - DECISION TREE PLOT

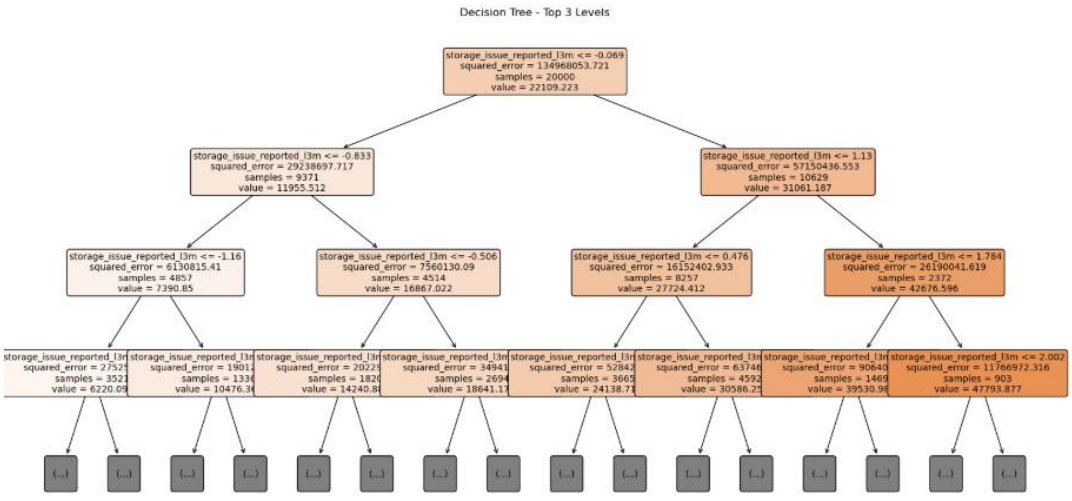


FIGURE 45 - DECISION TREE NODES AND LEVELS

4. SVR

SVR aims to find a linear function in a (possibly higher-dimensional) feature space that approximates the target values within a certain margin of tolerance (ϵ), while keeping the function as simple as possible (by minimizing the norm of the coefficients).

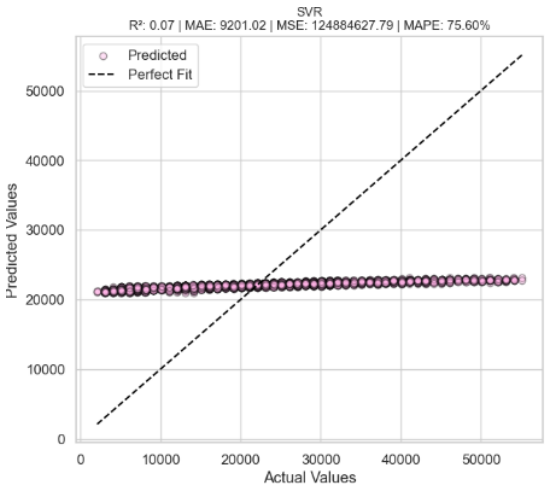


FIGURE 46 - SVR MODEL PLOT

- The scatter plot shows a loosely scattered pattern of predicted versus actual values, with a weaker alignment around the diagonal compared to previous models, indicating less precise predictions.
- The SVR model's error metrics are much higher, with an MAE of about 9201 and an MSE around 1.25×10^9, reflecting larger average and squared prediction errors.
- The very low R-squared (~0.067) reveals that the SVR explains only about 6.7% of the variance, indicating poor model fit and overall weak performance on this dataset.

Rishabh Gupta

5. Random Forest

An ensemble method that averages predictions from multiple decision trees

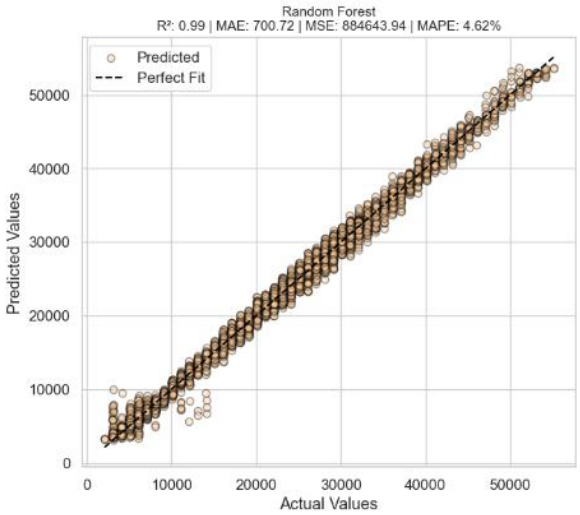


FIGURE 47 - RANDOM FOREST MODEL PLOT

- The scatter plot shows a very strong positive linear relationship between predicted and actual values, with points tightly clustered around the diagonal, indicating highly accurate predictions by the

Random Forest Regressor.

- The Mean Absolute Error (MAE) is approximately 697.73, the lowest among all models, reflecting the smallest average absolute prediction errors.
- The Mean Squared Error (MSE) is about 8.77 × 10^5, also the lowest observed, and the R-squared value is approximately 0.9934, indicating the model explains 99.34% of the variance and fits the data excellently.

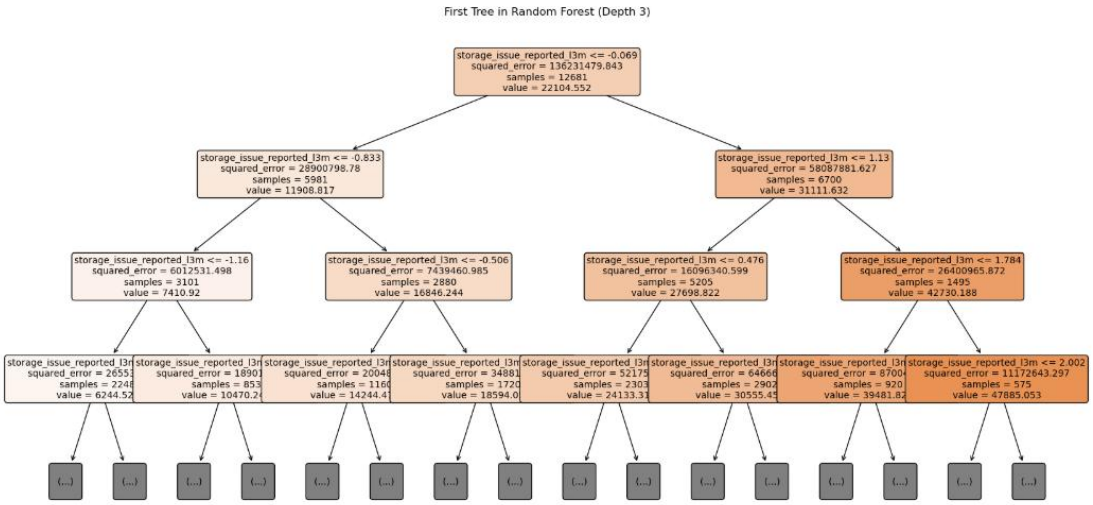


FIGURE 48 - RANDOM FOREST TREE PLOT

- The root node of this individual tree splits on storage_issue_reported_13m (<= -0.069), mirroring the single Decision Tree model and highlighting its key role.
- Subsequent splits in the tree continue to use storage_issue_reported_13m with varying thresholds, showing consistent feature importance.
- Squared error decreases as we move down the tree, indicating effective error reduction during tree construction.
- Leaf nodes at depth 3 show a range of predicted values, representing early-stage averaging that feeds into the overall Random Forest prediction.
- With a limited depth of 3, this tree captures broad patterns; the full Random Forest aggregates many such trees, enhancing accuracy and robustness.

The ensemble approach of the Random Forest, combining multiple trees, allows it to model complex relationships better than individual models, resulting in superior predictive performance and generalization.

Pishabh Gupta

6. Gradient Boosting

This is a powerful ensemble technique used for both regression and classification tasks.

- The scatter plot shows an exceptionally strong positive linear relationship between predicted and actual values, with points tightly clustered along the diagonal, indicating highly accurate predictions.
- The Gradient Boosting Regressor achieves the lowest errors so far, with a Mean Absolute Error (MAE) of approximately 689.49 and a Mean Squared Error (MSE) of about 8.36 x 10^5, reflecting minimal average and squared prediction errors.
- With an R-squared of approximately 0.9937, this model explains 99.37% of the variance, making it the best-performing model by effectively capturing complex data patterns through sequential learning.

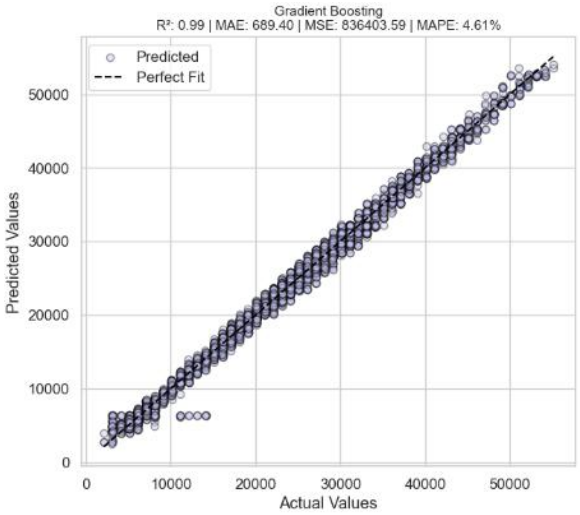


FIGURE 49 - GRADIENT BOOSTING MODEL PLOT

Let's analyze and interpret the performance Actual and Predicted values of all the regression models so far –

- Ensemble Models Win: Gradient Boosting and Random Forest outperformed others, achieving the lowest MAPE by effectively combining multiple learners.
- Linear Models Perform Decently: Linear and Ridge Regression fit reasonably well, showing a strong linear trend but lacking the flexibility of tree-based methods.
- Decision Tree and SVR Results: The Decision Tree improved on linear models by capturing non-linear patterns without severe overfitting, while SVR underperformed with default settings, showing very low R-squared and poor prediction trends.

Rishabh Gupta

Model Performance Comparison (R^2, MAE, MSE, MAPE)

Model	R^2 Score	MAE	MSE	MAPE (%)
Linear Regression	0.977	1303.975	3093785.29	9.066
Ridge Regression	0.977	1304.013	3093803.482	9.066
Decision Tree	0.987	883.128	1763136.218	5.79
Random Forest	0.993	700.72	884643.937	4.621
Gradient Boosting	0.994	689.397	836403.589	4.607
SVR	0.067	9201.024	124884627.786	75.595

TABLE 10 – MODEL PERFORMANCE COMPARISONS

In summary, Gradient Boosting Regressor is the most effective model here based on metrics like MAPE, but Random Forest offers a strong, simpler alternative, while SVR's poor performance highlights the need for careful hyperparameter tuning.

➢ Random Forest - Hyperparameter tuning

We plot Output Best Model – Refer code

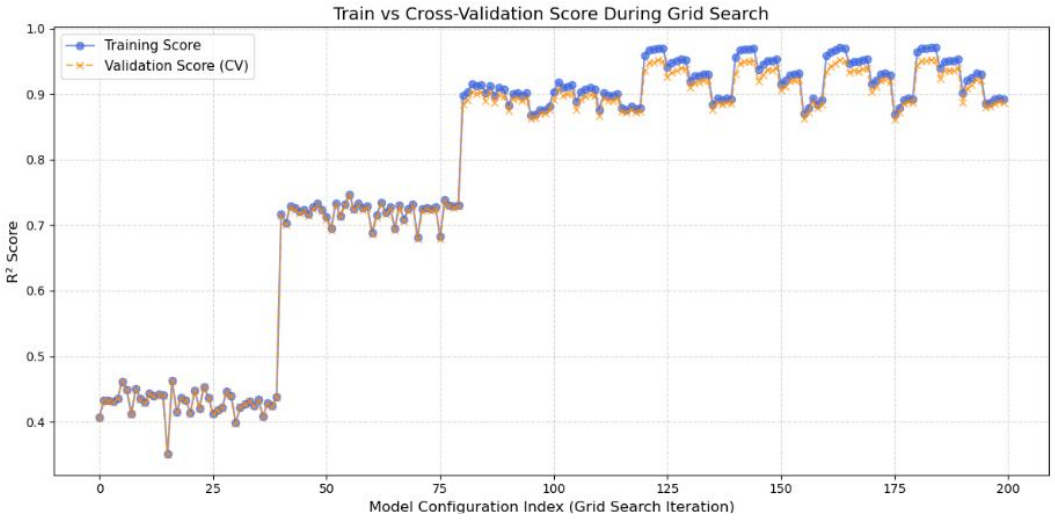


FIGURE 50 - GRID SEARCH TRAIN VS TEST - HYPERTUNED RF

The training and cross-validation scores are consistently close with high R^2 (up to ~0.95), indicating the model fits well without overfitting and generalizes effectively.

Let's examine how each hyperparameter in our grid_params impacts model performance on both training and validation sets using validation curves. The objectives are to identify:

- Which hyperparameters have the greatest effect on performance.
- Signs of overfitting or underfitting from trends and variability.

Rishabh Gupta

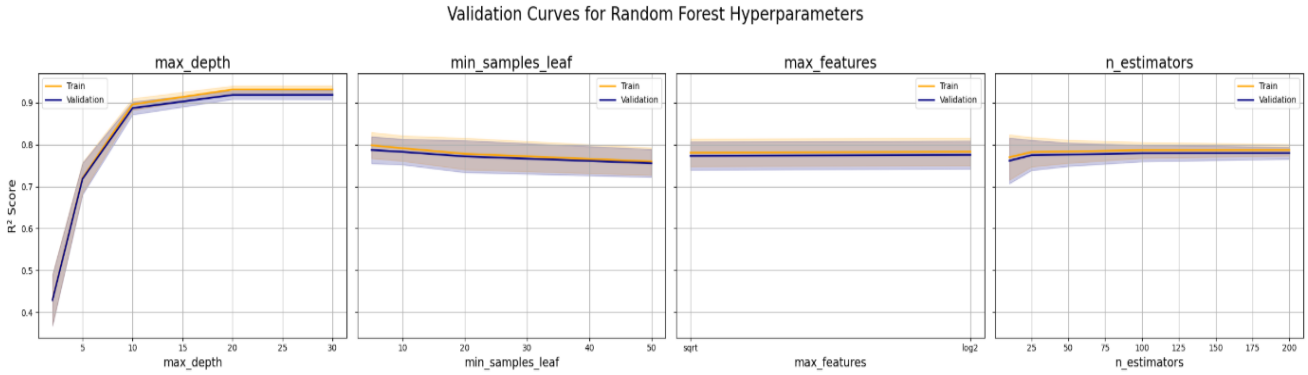


FIGURE 51 - VALIDATION CURVES FOR HYPERTUNED RANDOM FOREST

- max_depth: Validation score peaks around 10–20; deeper trees risk overfitting.
- min_samples_leaf: Values between 1 and 10 work best; higher values may cause underfitting.
- max_features: 'sqrt' performs slightly better and more consistently than 'log2'.
- n_estimators: Performance stabilizes around 100–150 trees.

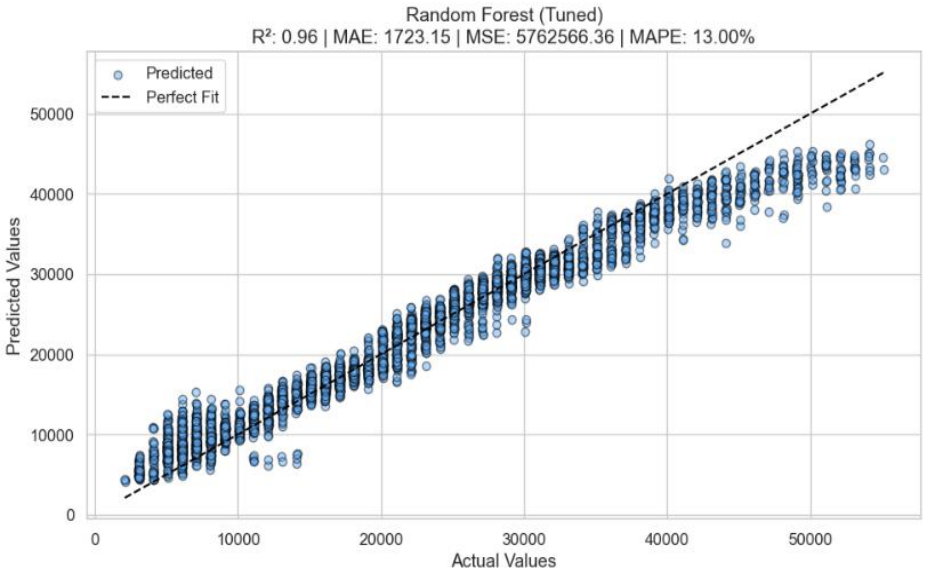


Figure 52 - Random Forest Tuned Model

There is a strong positive correlation between predicted and actual values with a high R^2 of 0.96 explaining most variance, an average MAE of ~1647 units, MSE around 5.3 million due to larger errors, and tuning slightly lowered R^2 while increasing error metrics for better generalization.

Also lets observe-

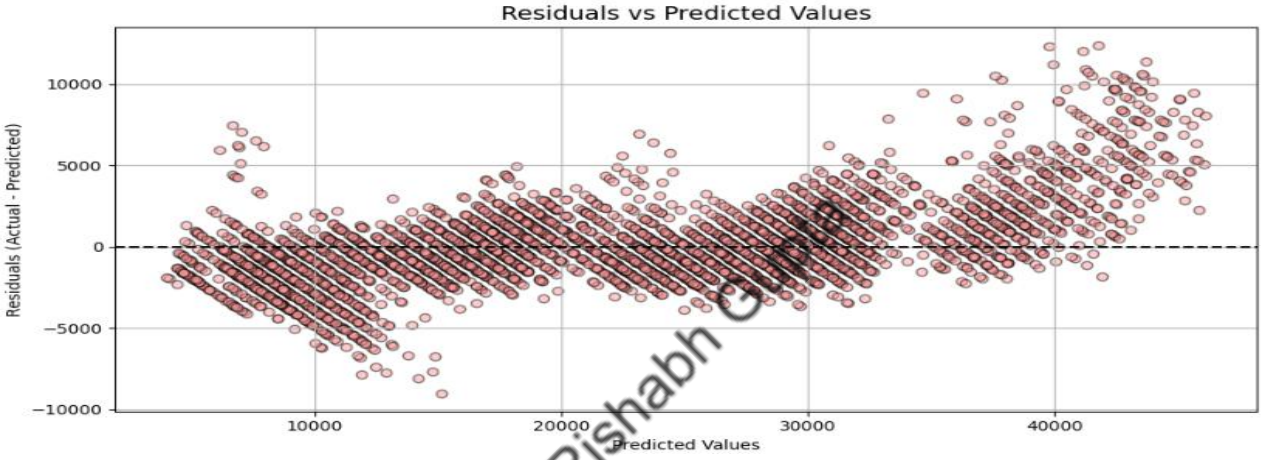


FIGURE 53 - RF TUNED - RESIDUALS PLOT

- Residuals show increasing spread with higher predicted values, indicating heteroscedasticity.
- A curved pattern in residuals suggests the model may miss some non-linear relationships.
- No clear outliers, but residuals are unevenly distributed around zero.

Conclusion: The Random Forest model performs robustly across hyperparameters, with the final chosen configuration being both accurate and reliable.

➢ Gradient Boosting - Hyperparameter tuning

Then plotting output set.

Training R-squared remains consistently high, while validation scores fluctuate widely, indicating likely overfitting for many hyperparameter settings. However, validation scores plateau near a high value toward the end, showing grid search effectively identified better generalizing hyperparameters.

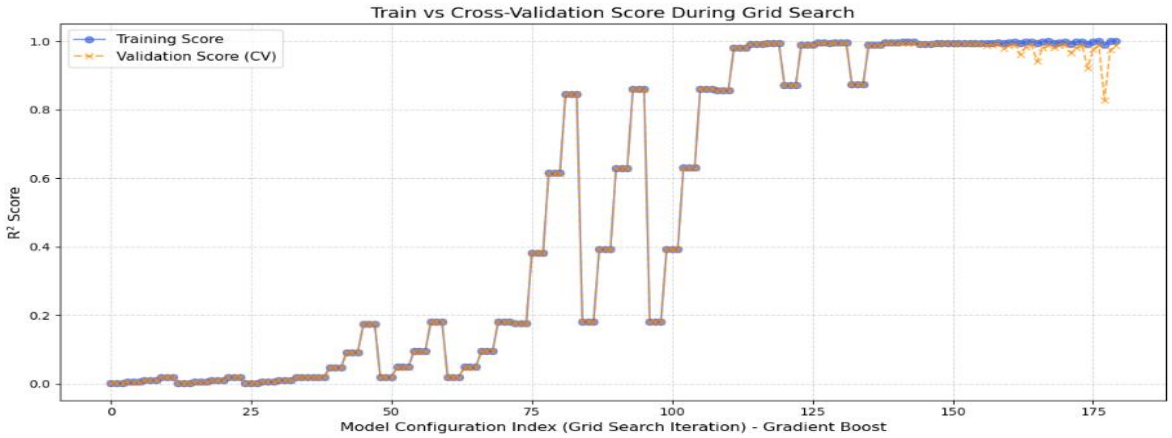


FIGURE 54 - GRID SERACH TRAIN VS TEST - HYPERTUNED GB

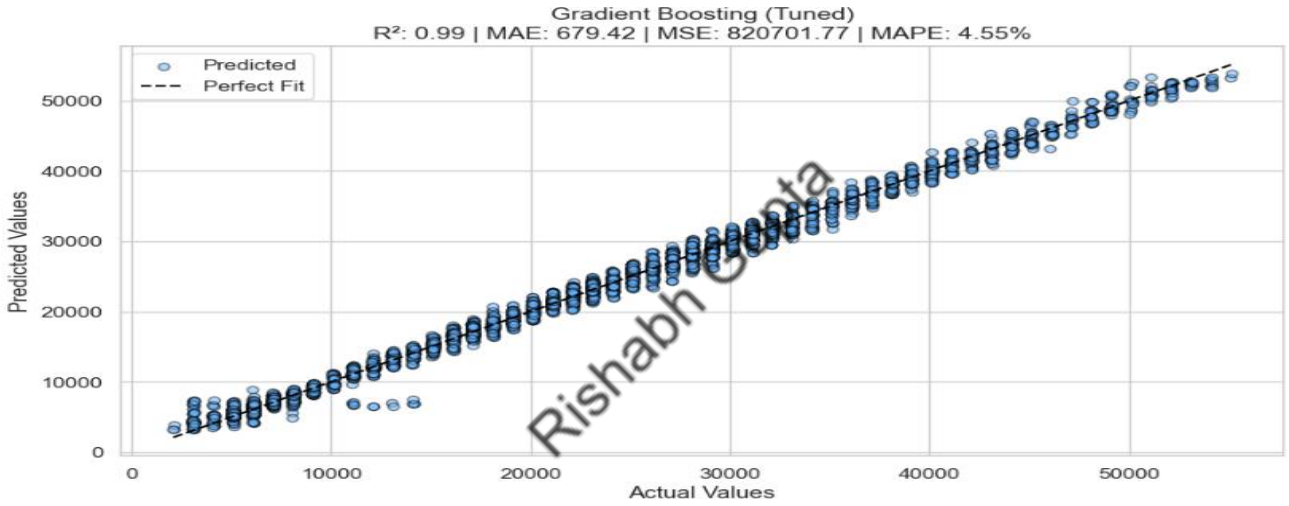


FIGURE 55 – GRADIENT BOOST PLOT

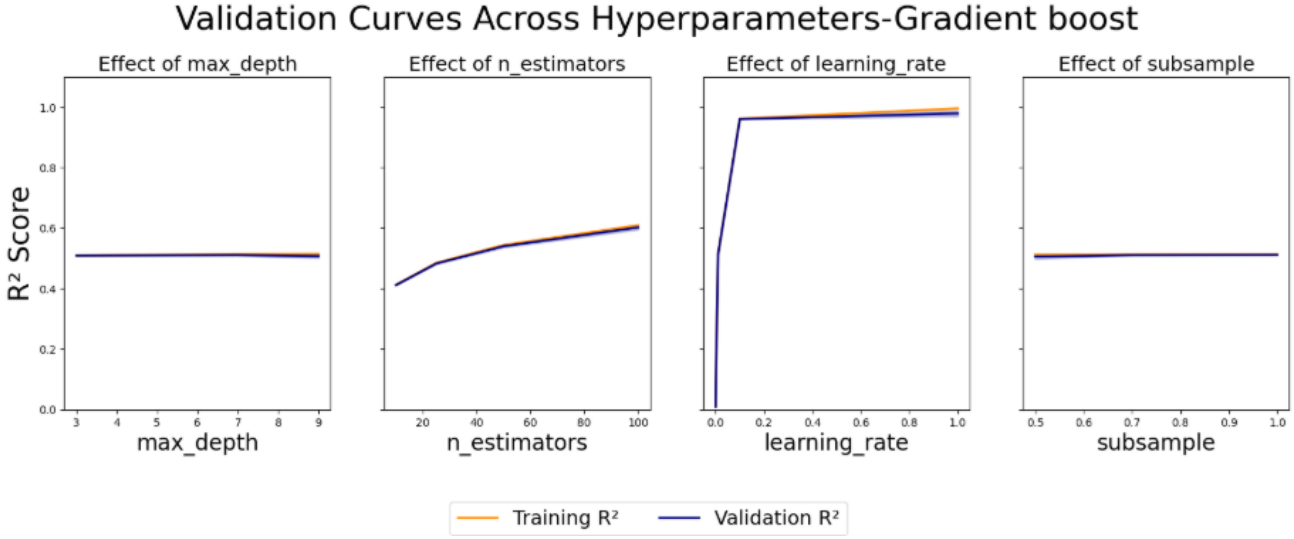


FIGURE 56 - VALIDATION CURVES FOR HYPERTUNED GRADIENT BOOST MODEL

- max_depth Insensitivity: R^2 scores remain stable across tested depths, indicating little impact on performance within this range.
- n_estimators Improvement: Increasing estimators steadily boosts training and validation R^2, with gains plateauing at higher counts.
- learning_rate Sensitivity: Learning rate strongly affects performance—too low (0.01) hurts accuracy, while 0.1 yields a sharp improvement before leveling off.

With an R^2 of 0.99, MAE of 679.42, and MAPE of 4.55%, the model delivers highly accurate, consistent predictions closely aligned with actual values, demonstrating strong generalization and superior visual fit.

➢ Ensembled (Gradient + RF)

Please refer code for execution.

- Strong Positive Correlation: The top scatter plot shows predicted values tightly clustered around the perfect fit line, indicating a strong correlation with actual values.
- High R-squared (0.99): The model explains 99% of the variance in the test set, demonstrating excellent predictive power.
- Error Metrics: The MAE is approximately 1060 units, and the MSE is about 1,987,914, reflecting average prediction errors and their squared values.
- Residuals & Comparison: Residuals are mostly randomly distributed with slight heteroscedasticity and subtle non-linearity; compared to the tuned Gradient Boosting model, this ensemble has slightly higher errors but may offer more stable predictions.

Rishabh Gupta

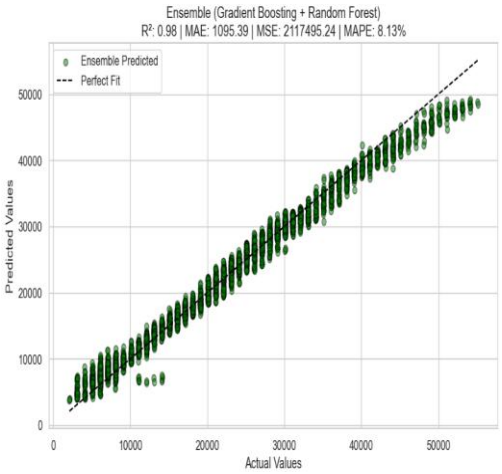


FIGURE 57 - ENSEMBLE (GRADIENT + RF) MODEL

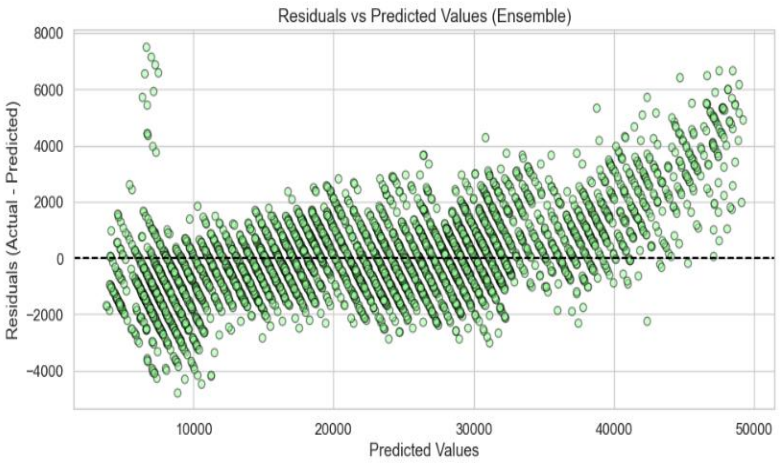


FIGURE 58 - ENSEMBLE MODEL RESIDUAL PLOT

Final Model Comparison

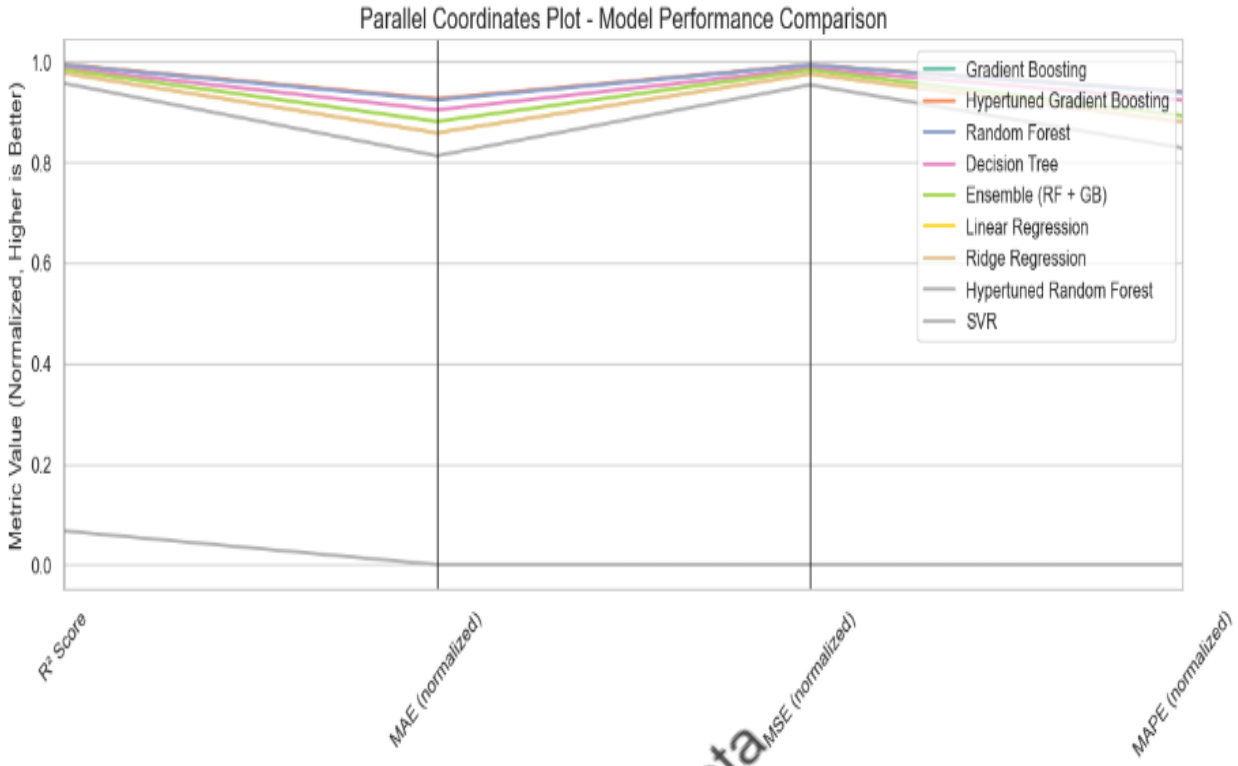


Figure 60 - Parallel coordinates plot for Model performances

Rishabh Gupta

- Top Performers with Low MAPE: Gradient Boosting, Hypertuned Gradient Boosting, and Random Forest lead the pack with R^2 values near 1 and low error metrics; all have MAPE under 5% (GB: 4.607%, Hypertuned GB: 4.545%, RF: 4.621%), indicating excellent accuracy.

- Ensemble Performance Anomaly: The RF + GB ensemble, despite a strong R^2 of 0.984, shows higher error metrics and a notably higher MAPE of 8.13%, suggesting this combination didn't improve accuracy over individual models.

- Hypertuning Effects: Hypertuned Gradient Boosting slightly improves accuracy (MAPE 4.545% vs. 4.607%), while Hypertuned Random Forest performs worse than the original, with a much higher MAPE of 12.996%, indicating suboptimal tuning for RF.

- Other Models and Trends: Linear and Ridge Regression have moderate performance with MAPE around 9%, while Decision Tree performs better (R^2 0.987, MAPE 5.79%). SVR severely underperforms (MAPE 75.6%), highlighting the need for proper hyperparameter tuning. Error metrics generally align, and R^2 remains the key differentiator across models.

Model Performance Comparison (Sorted by R^2 Score)

Model	R^2 Score	MAE	MSE	MAPE (%)
Gradient Boosting	0.994	689.397	836403.589	4.607
Hypertuned Gradient Boosting	0.994	679.42	820701.772	4.545
Random Forest	0.993	700.72	884643.937	4.621
Decision Tree	0.987	883.128	1763136.218	5.79
Ensemble (RF + GB)	0.984	1095.387	2117495.239	8.13
Linear Regression	0.977	1303.975	3093785.29	9.066
Ridge Regression	0.977	1304.013	3093803.482	9.066
Hypertuned Random Forest	0.957	1723.153	5762566.36	12.996
SVR	0.067	9201.024	124884627.786	75.595

TABLE 11 – FINAL MODEL PERFORMANCE COMPARISONS

Hypertuned Gradient Boosting demonstrate the best overall performance **and our Final Model.**

- Highest R^2 (0.994), explaining 99.4% of variance
- Lowest MAE (679.42), indicating minimal average error
- Lowest MSE (820,701), showing fewer large errors
- Lowest MAPE (4.545%), meaning predictions are off by just ~4.5% on average

Rishabh Gupta

- **Vanilla Gradient Boosting** performs nearly as well; **Random Forest** is close but slightly behind in all errors.
- The **Ensemble model** shows promising results but doesn't surpass the hypertuned GB.
- **SVR** performs poorly with R^2 of 0.067 and much higher errors.

Conclusion: Based on all key metrics, hypertuned Gradient Boosting offers the best accuracy and reliability for this dataset.

Insights and Recommendations

Business Impact of the Optimum Model – Hypertuned Gradient Boosting

- Superior Accuracy & Reliability: Delivers precise forecasts for key metrics (sales, risk, demand), reducing costly errors and enabling confident decision-making.
- Smarter Strategic Decisions: Low MAE and MAPE support optimized inventory, budgeting, marketing, and resource allocation, minimizing uncertainty and operational risks.
- Operational Efficiency Gains: Accurate predictions improve supply chain management, production planning, and logistics, leading to tangible cost savings.
- Competitive Edge: Outperforms traditional models, allowing faster, data-driven responses to market changes and customer needs.

Expected Business Outcomes

- Better Decisions: Accurate forecasts improve planning and financial control needed.
- Higher Efficiency: Enables just-in-time inventory and predictive maintenance.
- Improved Customer Retention: Forecasts support personalized engagement.
- Revenue Growth: Informs pricing, promotions, and cross-selling opportunities.
- Risk Reduction: Accurately flags high-risk cases, minimizing losses.

Rishabh Gupta

Our goal was to optimize product shipment by using data-driven prediction models. Through extensive analysis, we found that retail presence and refill requests are strong indicators of demand, suggesting that warehouses serving more retail shops or with frequent refills should be prioritized for forecasting and optimized replenishment schedules.

Boost and follow four warehouse types:

- High-demand hubs → Need frequent, larger shipments.
- Efficient operators → Should be used as performance benchmarks.
- Remote locations → Use bulk logistics to reduce cost.
- Problematic warehouses → Need operational audits & improvements to resolves issues

To summarize:

- Focus forecasting efforts on retail-heavy, refill-active zones.
- Improve infrastructure in high-volume warehouses.
- Explore growth opportunities in the East zone.
- Use top-performing model to guide strategic supply decisions at large scale.

Together, these actions can improve customer satisfaction, boost revenue, reduce costs, and create a more agile and responsive supply chain."

Appendix

List of Tables

Rishabh Gupta

List of Figures

Rishabh Gupta

Rishabh Gupta