Supply Chain Management

Capstone project

FASNA

11/08/24

CONTENTS:

TABLES:

FIGURES:

Business Problem:

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.

Goal & Objective:

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.

File: Data.csv Target variable: product_wg_ton

Data Dictionary

Ware_house_ID : Unique Warehouse id where product is prepared for dispatch

WH_Manager_ID : Manager Id present in the warehouse

zone : Zone of the Warehouse

WH_regional_zone : Regional Zone of the warehouse

num_refill_req_l3m : Refilling request received by the warehouse in the last 3 months

transport_issue_l1y : No. of transport issued for warehouse in last 1 year

Competitor_in_mkt : No. of competitors in the market

retail_shop_num : Number of retail shops who sell noodlesproduced by the warehouse

wh_owner_type : The warehouse is owned by the company or it is on rent

distributor_num : No. of distributor who works between warehouse and retail shops

flood_impacted : Is the warehouse in a flood impacted area or not

flood_proof : Warehouse is having flood proof indicator

electric_supply : Does the warehouse have proper electric supply along with some power backup

dist_from_hub : distance from the warehouse to production hub

workers_num : no. workers in the warehouse

wh_est_year : warehouse establishment year

storage_issue_reported_l3m : storage issues reported by the warehouse in the last 3 months

govt_check_l3m : Government checking in last 3 months

temp_reg_mach : warehouse having temperature regulating machine indicator or not

approved_wh_govt_certificate : Type of approval warehouse having been issued by government

wh_breakdown_l3m : Number of times the warehouse faces thebreakdown in the last 3 months

product_wg_ton : Product weight

Objective Clarification of problem statement

Need of the Study/Project

The study aims to address the following needs:

- **Reduce Inventory Costs**: Minimize excess inventory and reduce associated holding costs.
- **Improve Service Levels**: Ensure timely availability of products in high-demand areas, leading to better customer satisfaction.
- **Enhance Operational Efficiency**: Streamline logistics and distribution processes, reducing transportation and handling costs.
- **Increase Market Share**: By meeting demand accurately, the company can capture a larger market share and reduce the risk of stockouts.
- **Inform Strategic Decisions**: Insights from demand patterns can inform targeted marketing campaigns and strategic decisions.

Understanding Business/Social Opportunity

Optimizing the supply chain will provide multiple benefits:

Significance of Demand Planning and Supply Chain Management

- **Market Responsiveness**: Efficient demand planning allows companies to respond quickly to market changes, ensuring products are available where and when they are needed.
- **Cost Efficiency**: Optimized supply chains reduce costs related to storage, transportation, and handling, directly impacting the bottom line.
- **Sustainability**: Reducing excess inventory and transportation needs contributes to a more sustainable and environmentally friendly operation.
- **Customer Satisfaction**: Ensuring product availability enhances customer satisfaction and loyalty, which are crucial for business growth.

Optimization Techniques with Facts & Figures

Statistical Forecasting: Techniques such as time series analysis, regression models, and machine learning algorithms can predict future demand based on historical data. For example, companies like Walmart use advanced forecasting models to reduce inventory costs by 10-15%.

Inventory Optimization: Tools like Economic Order Quantity (EOQ), Just-In-Time (JIT), and ABC analysis help in maintaining optimal inventory levels. Implementing JIT, Toyota managed to reduce its inventory costs by 20%.

Demand Sensing: Real-time data analytics allows companies to sense demand shifts quickly. Procter & Gamble uses demand sensing to reduce forecast errors by 30-40%, resulting in better alignment of supply with actual demand.

Supply Chain Coordination: Collaborative planning and supply chain visibility enhance coordination among suppliers, manufacturers, and distributors. Companies like Dell use supply chain coordination to reduce lead times and increase efficiency.

Network Optimization: Analyzing and optimizing the distribution network ensures products are stored and transported efficiently. Coca-Cola's network optimization initiatives have led to a 5-10% reduction in logistics costs.

Conclusion

Optimizing the supply chain and demand planning for the FMCG company's instant noodles business will lead to significant cost savings, improved service levels, and strategic market advantages. By leveraging historical data and advanced analytics, the company can ensure a balanced supply across all warehouses, aligning with actual demand and driving overall business success.

Dataset Comprehension-

	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_num	...	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_wg_ton
0	WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	2	4651	...	1	91	29.0	NaN	13	0	A	5	15	17115
1	WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	4	6217	...	1	210	31.0	NaN	4	0	A	3	17	5074

	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_num	...	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_wg_ton
2	WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	4	4306	...	0	161	37.0	NaN	17	0	A	6	22	23137
3	WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	2	6000	...	0	103	21.0	NaN	17	1	A+	3	27	22115
4	WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	2	4740	...	1	112	25.0	2009.0	18	0	C	6	24	24071

Table-1-Head of te dataset

Understanding How Data Was Collected

Time Frame: The data spans the last 3 months for refilling requests and storage issues, and the last 1 year for transport issues. The exact date range is not specified but can be inferred from these variables.

Frequency: The data appears to be collected at regular intervals, considering the presence of metrics for specific periods (e.g., last 3 months, last 1 year).

Methodology: The data likely comes from the company's internal logistics and supply chain management systems, with metrics on warehouse operations, transportation, retail distribution, and environmental factors.

Data set have - 25000 rows ,24 columns----------

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	workers_num	wh_est_year	storage_issue_reported_l3m	temp_reg_mach	wh_breakdown_l3m	govt_check_l3m	product_wg_ton
count	25000.00	2500 0.00	25000.00	25000.00	25000.00	2500 0.0	2500 0.00	2500 0.00	250 00.00	240 10.00	1311 9.00	25000.00	250 00.00	25000.00	250 00.00	2500 0.00

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	workers_num	wh_est_year	storage_issue_reported_l3m	temp_reg_mach	wh_breakdown_l3m	govt_check_l3m	product_wg_ton
mean	4.09	0.77	3.10	4985.71	42.42	0.1	0.05	0.66	163.54	28.94	2009.38	17.13	0.30	3.48	18.81	22102.63
std	2.61	1.20	1.14	1052.83	16.06	0.3	0.23	0.47	62.72	7.87	7.53	9.16	0.46	1.69	8.63	11607.76
min	0.00	0.00	0.00	1821.00	15.00	0.0	0.00	0.00	55.00	10.00	1996.00	0.00	0.00	0.00	1.00	2065.00
25%	2.00	0.00	2.00	4313.00	29.00	0.0	0.00	0.00	109.00	24.00	2003.00	10.00	0.00	2.00	11.00	13059.00
50%	4.00	0.00	3.00	4859.00	42.00	0.0	0.00	1.00	164.00	28.00	2009.00	18.00	0.00	3.00	21.00	22101.00
75%	6.00	1.00	4.00	5500.00	56.00	0.0	0.00	1.00	218.00	33.00	2016.00	24.00	1.00	5.00	26.00	30103.00
max	8.00	5.00	12.00	11008.00	70.00	1.0	1.00	1.00	271.00	98.00	2023.00	39.00	1.00	6.00	32.00	55151.00

Table-2-Descriptive summary

Information about Data

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----
```

Data visualizations

UNIVARIATE ANALYSIS

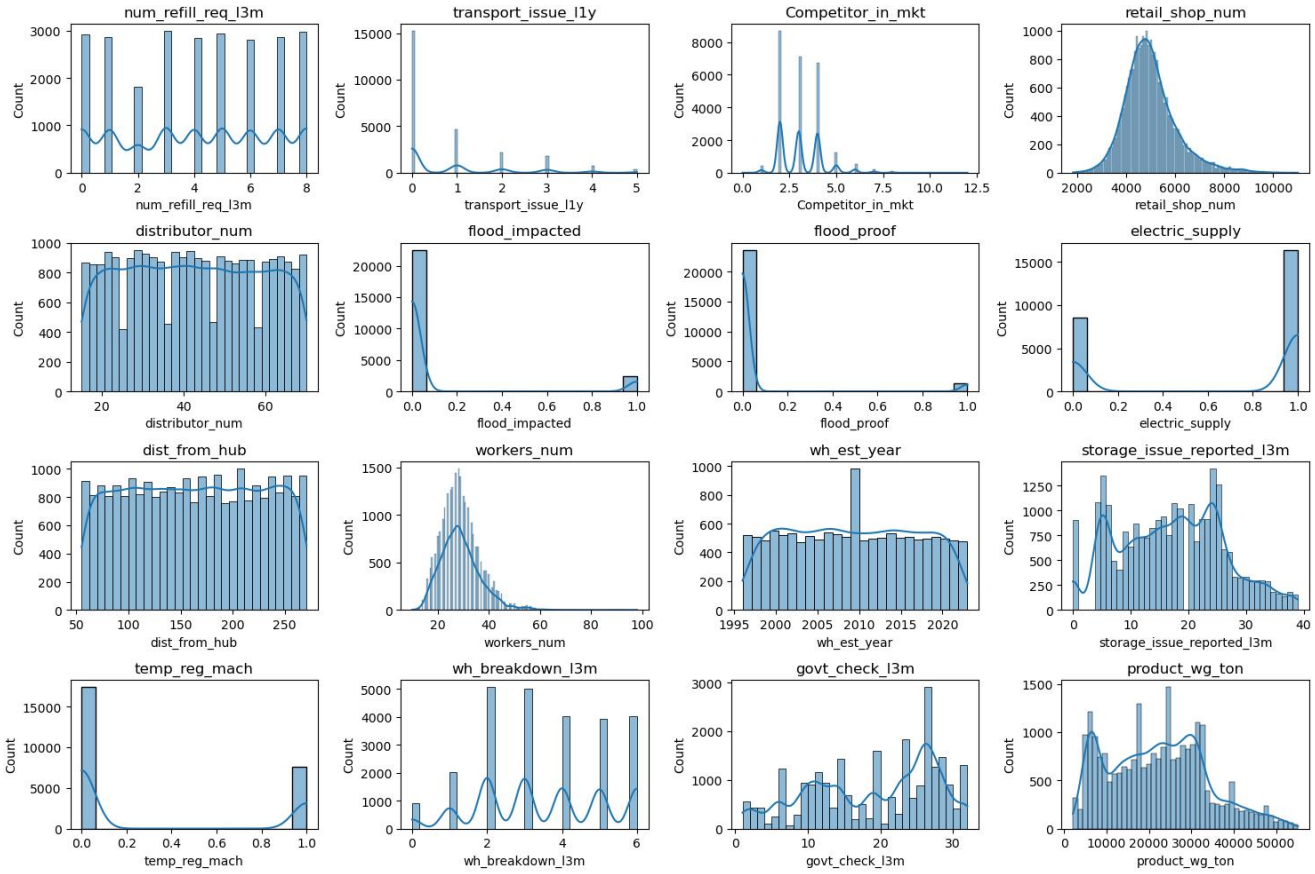


Figure-1-Histogram of all numeric variables

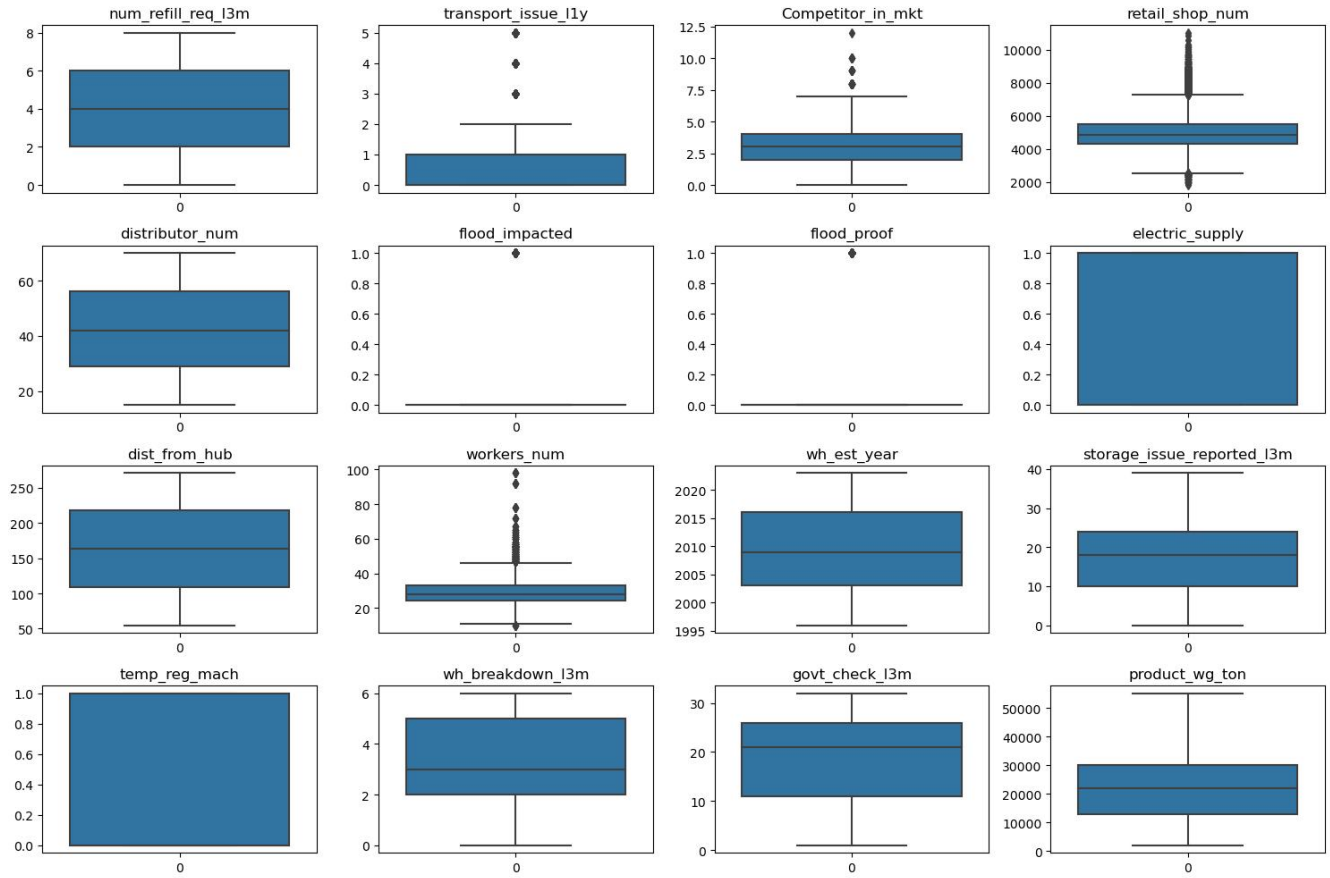


Figure-2-boxplots

we can see the target variable product_wg_ton is approximatly normal

Number of Refill Requests (num_refill_req):
The data shows a peak around 3,000 requests.
There seems to be a consistent trend over time.
Transportation Issues (transport_issue_by):
The graph indicates that transportation issues are reported.
Competitor in the Market (Competitor_in_mkt):
The data suggests the presence of competitors.
Retail Shop Numbers (retail_shop_num):
The graph displays varying numbers of retail shops.
It's essential to consider the time frame and location.
Distribution from Hub (dist_from_hub):
The data shows a range of values.
Understanding the context is crucial for interpretation.
Storage Issues Reported (storage_issue_reported):
The graph highlights storage-related challenges.
Investigating the causes and solutions is necessary.

distribution of data in categories for categorical ones

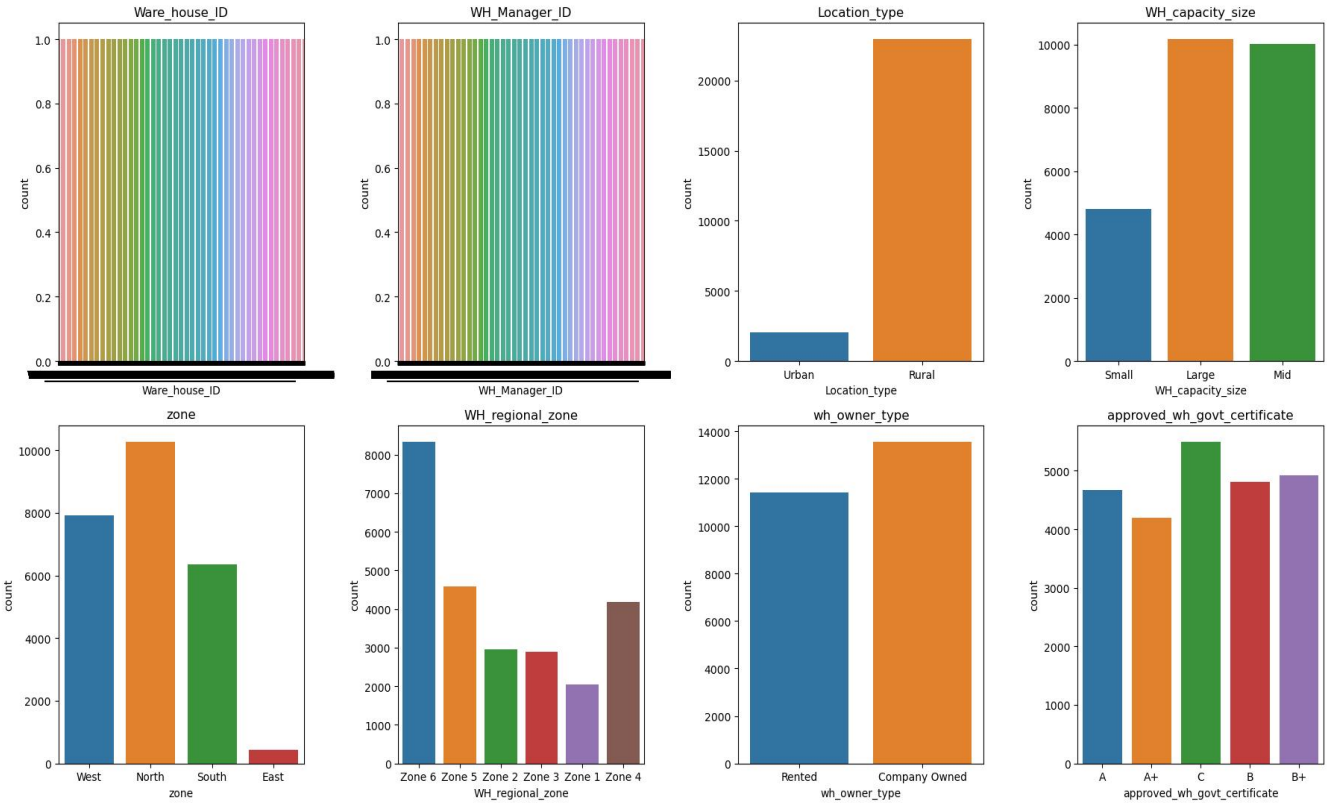


Figure-3-countplot of all categorical variables

Warehouse ID (Ware_house_ID):

The distribution across different warehouse IDs appears relatively uniform, with only minor variations.
No specific warehouse ID dominates significantly, which suggests a balanced distribution of products across warehouses.

Warehouse Manager ID (WM_Manager_ID):

Similar to warehouse IDs, the manager IDs also show a fairly uniform distribution.
No specific manager seems to handle a disproportionately large number of warehouses.

Location Type:

The majority of warehouses fall into the 'rural' category, indicating that urban areas have more warehouses.
There are fewer warehouses in urban areas, which might impact supply and demand dynamics differently.

Warehouse Capacity Size (WH_capacity_size):

The 'large' capacity category has the highest count, followed by 'mild' and 'small.'
This distribution could influence inventory management and logistics planning.

Regional Zones (WH_regional_zone):

The counts vary significantly across different Zones.
The highest count is in Zone 6, while the Zone 1 has the lowest count.
This information can guide targeted marketing efforts based on regional demand.

Zones

The counts vary significantly across different regions.
The highest count is in the North region, while the East region has the lowest count.
This information can guide targeted marketing efforts based on regional demand.

Ownership Type (wh_owner_type):

Most warehouses are either 'Rented' or 'Company Owned.'
The 'Company Owned' category has a higher count, suggesting a mix of ownership models.

'approved_wh_govt_certificate'

there are 5 types of certificates (A,A+, B,B+, C), with C_certification having the highest count.

Bivariate analysis

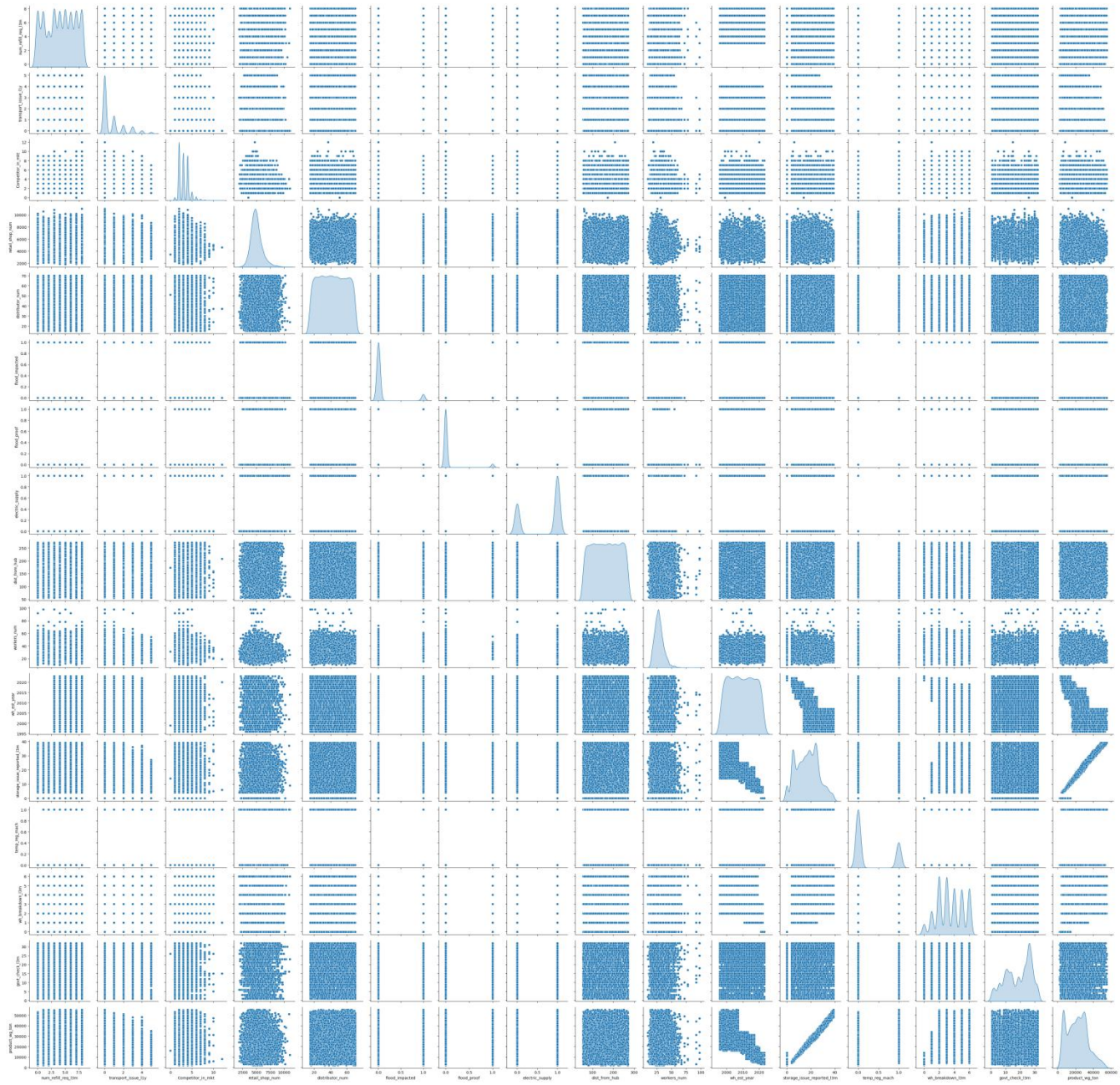


Figure-4-pair plot

Warehouse location type

	Location_type	product_wg_ton
0	Rural	501482582
1	Urban	51083241

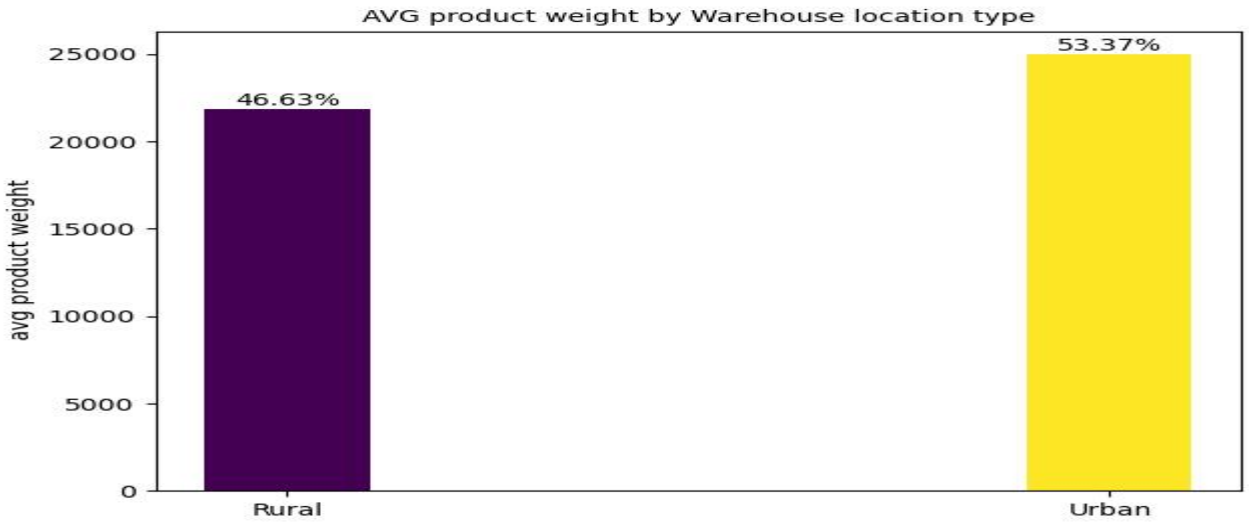


Figure-5-AVG product weight by Warehouse location type

- Urban areas receive 9.24% of the total product weight.
- Rural areas receive 90.76% of the total product weight.
- The product weight distributed to rural areas is significantly higher than that to urban areas.

Warehouse Owner Type

	wh_owner_type	product_wg_ton
0	Company Owned	299270114
1	Rented	253295709

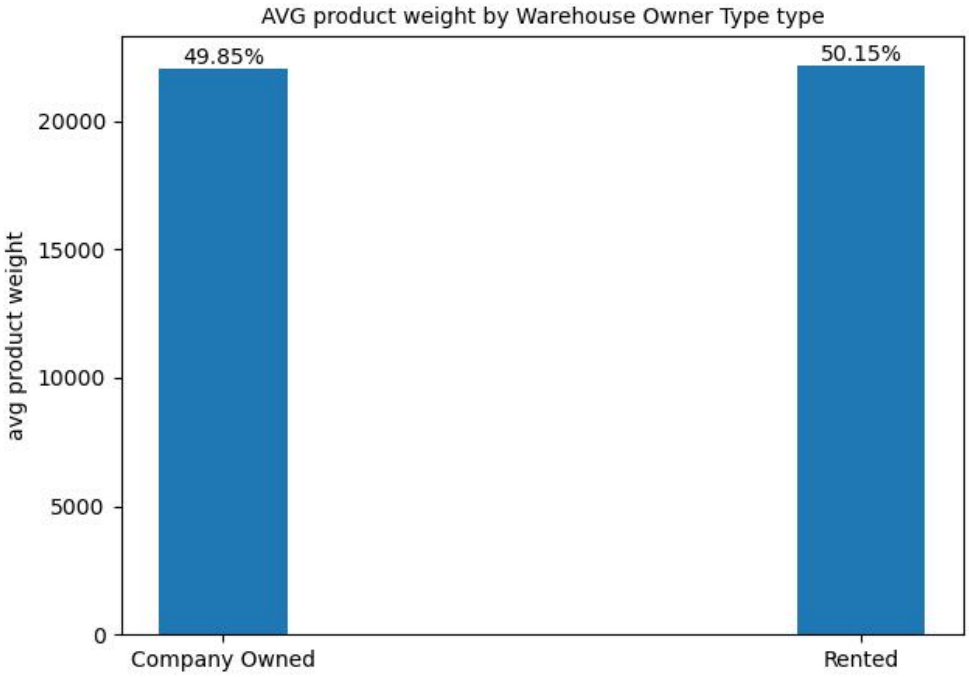


Figure-6-AVG product weight by Warehouse Owner Type type

- 46% of warehouses are rented, while 54% are company-owned.
- The average product weight in rented and company-owned warehouses is almost the same.

- Despite being fewer in number, company-owned warehouses handle a slightly higher total product weight than rented warehouses.

Temperature regulating machine availability

	temp_reg_mach	product_wg_ton
0	0	371425974
1	1	181139849

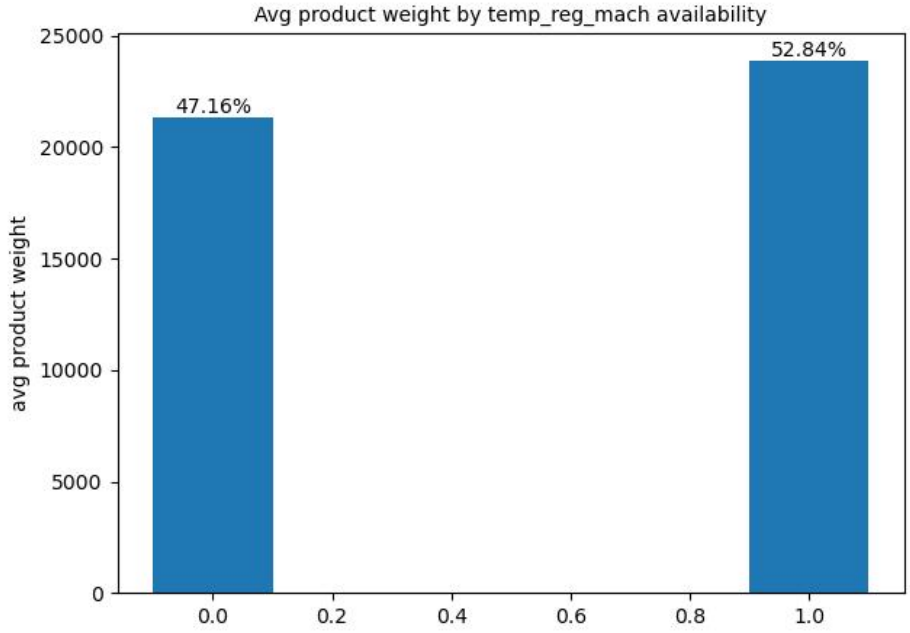


Figure-7-Avg product weight by temp_reg_mach availability

- 67.8% of warehouses do not have temperature regulating machines, while 32.7% do.
- Warehouses with temperature regulating machines handle a total product weight of 101,599,625 units, whereas those without handle 208,418,842 units.
- On average, warehouses without temperature regulating machines handle less product weight compared to those with this equipment available.

electric_supply

	electric_supply	product_wg_ton
1	1	362671887
0	0	189893936

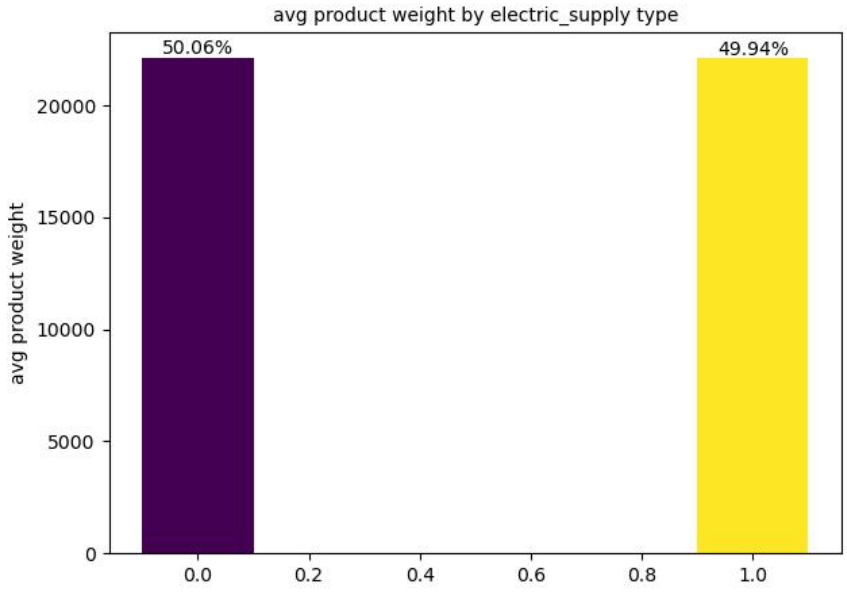


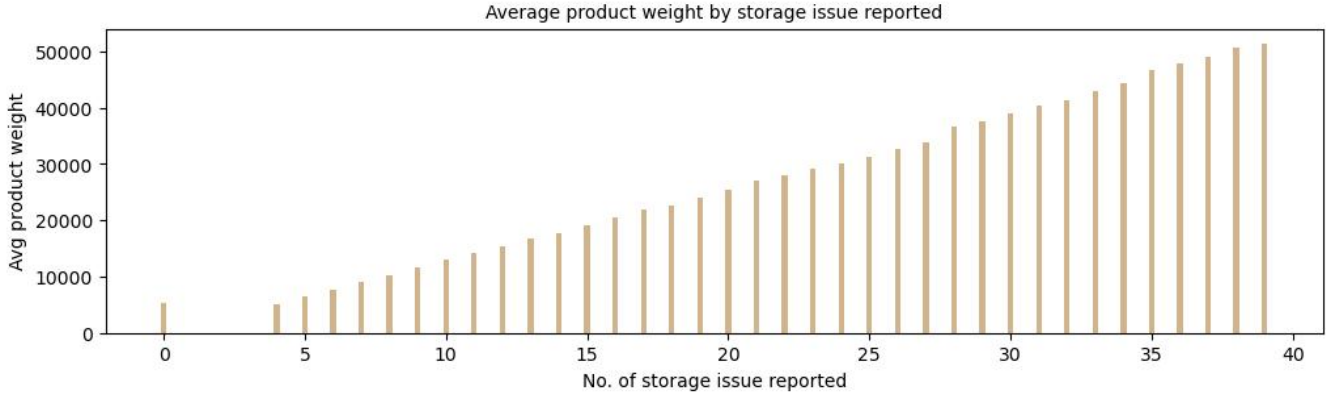
Figure-8-avg product weight by electric_supply type

- 65.63% of surveyed warehouses do not have accessible electric supply, while 34.37% do.
- Warehouses with accessible electric supply hold a total product weight of 203,327,830 units, whereas those without hold 106,690,637 units.
- On average, warehouses without accessible electric supply manage a lower product weight compared to those with electricity available.

Storage issue reported in the last 3 months

Unique values in Storage issue reported [13 4 17 18 23 24 6 11 22 9 29 19 14 28 25 12 8 0 34 16 38 21 15 36
31 20 10 32 27 26 7 37 30 5 39 33 35]

	storage_issue_reported_l3m	product_wg_ton
21	24	42904667
22	25	39461458
17	20	27006058
20	23	26797528
19	22	25472459



- **Figure-9-Average product weight by storage issue reported**

There is a strong correlation between product weight and storage issues, indicating a nearly linear relationship.

- As product weight increases, the occurrence of storage issues also escalates accordingly.

No. of warehouse breakdown in the last 3 months¶
Unique values in no. of warehouse breakdown [5 3 6 4 2 1 0]

	wh_breakdown_l3m	product_wg_ton
3	3	110723620
2	2	109370976
4	4	103310885
6	6	101532553
5	5	99024135

Figure-10-Average product weight by no of warehouse breakdown

- There is a positive correlation between product weight and the frequency of warehouse breakdowns, suggesting that as product weight increases, so does the likelihood of warehouse breakdowns occurring.
- Warehouses with the highest frequency of breakdowns also tend to handle higher total product weights.

ZONES
Unique values in zone ['West' 'North' 'South' 'East']

Unique values in zone ['West' 'North' 'South' 'East']

	zone	product_wg_ton
1	North	228165823
3	West	175111596
2	South	139540901
0	East	9747503

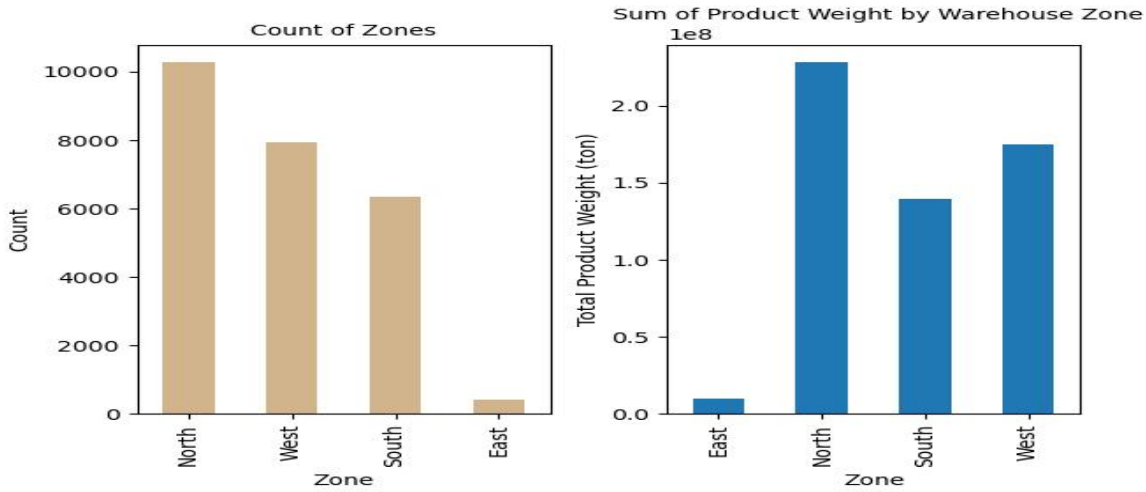


Figure-11-Sum of Product Weight by Warehouse Zone

- The North Zone has the highest production levels, while the East Zone has the lowest.
- The following chart shows that product weight remains fairly stable across different zones.
- Zones with more frequent product weight values have higher total product weights.

regional zone

Unique values in zone ['Zone 6' 'Zone 5' 'Zone 2' 'Zone 3' 'Zone 1' 'Zone 4']

	WH_regional_zone	product_wg_ton
5	Zone 6	184421651
4	Zone 5	101017613
3	Zone 4	92596029
1	Zone 2	66580768
2	Zone 3	63290230
0	Zone 1	44659532

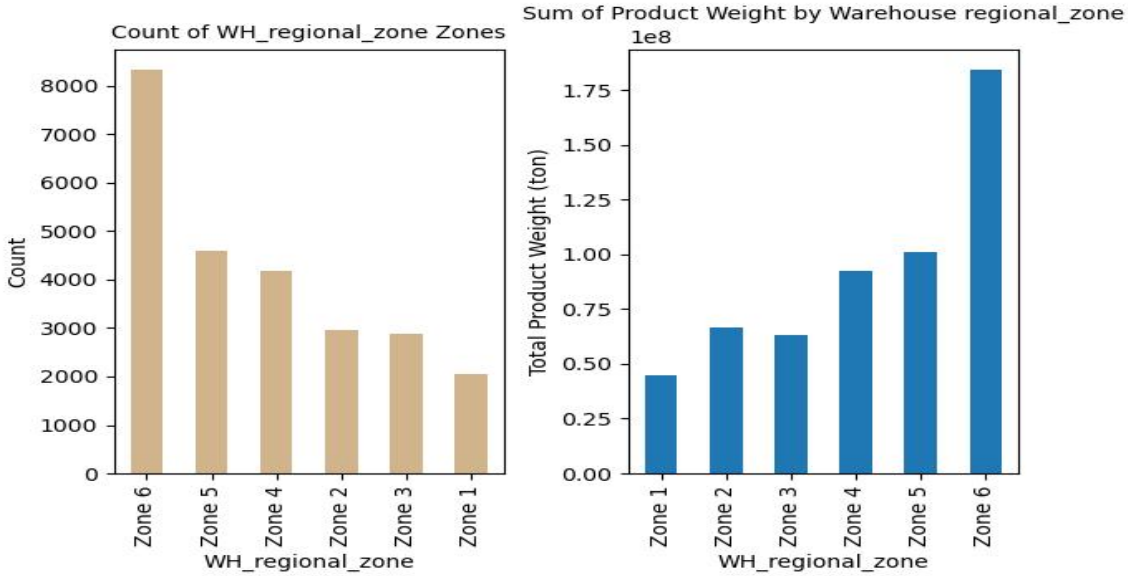


Figure-12-Sum of Product Weight by Warehouse regional_zone

- Regional Zone 6 has a notably high number of warehouses.

- The accompanying chart indicates that product weight remains relatively consistent across different Regional Zones.
- Higher total product weight is observed in zones where the frequency of product weight values is greater.

Government certification grades

Unique values in Government certification grades ['A' 'A+' 'C' 'B' 'B+' nan]

	approved_wh_govt_certificate	product_wg_ton
4	C	115184830
0	A	112676348
1	A+	111974920
3	B+	105499193
2	B	102299663

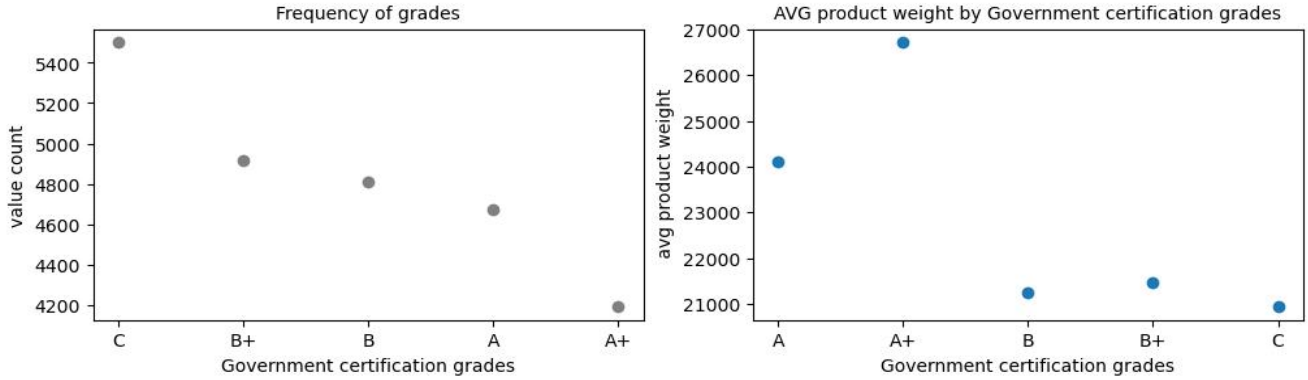


Figure-13-AVG product weight by Government certification grades

- The average product weight across different grades shows that warehouses with the highest government approvals also have the highest product weights.
- Despite being the lowest designation, C grade warehouses have the highest total product weights, mainly because most warehouses fall under the C grade category.

govt checking in last 3 months

Unique values in govt checking in last 3 months [15 17 22 27 24 3 6 2 28 1 11 9 12 21 19 8
14 23 26 29 10 13 30 32
7 25 31 20 5 16 18 4]

	govt_check_l3m	product_wg_ton
25	26	64681731
22	23	41001490
18	19	35595929
13	14	31892835
27	28	31869593

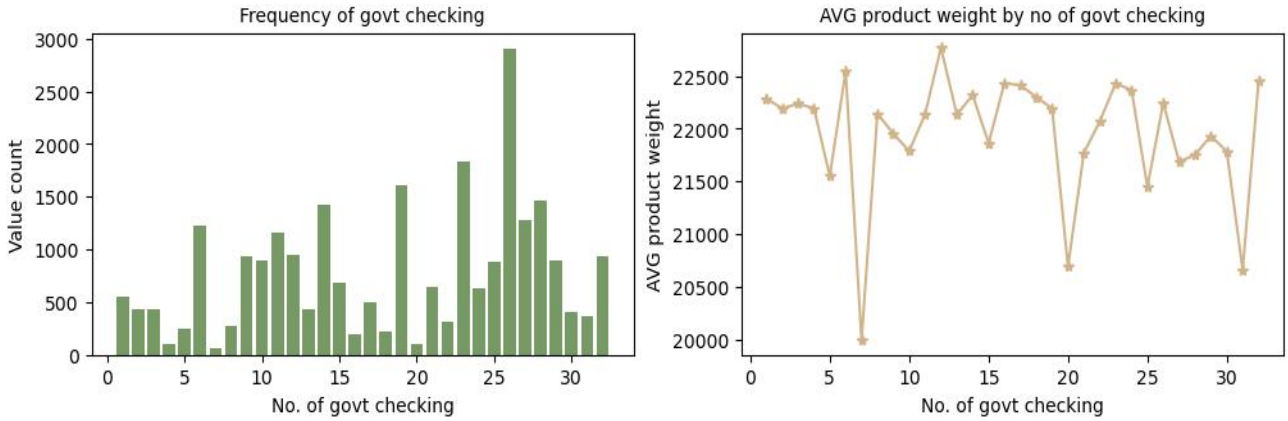


Figure-14-AVG product weight by no of govt checking

- The number of government checks in the last 3 months ranges from 1 to 32 times, with the bar graph showing significant variations in the frequency of these checks.
- As the frequency of government checks increases, the total product weight also increases. However, the average product weight remains relatively stable.

workers_num

Unique values in No. of workers [29. 31. 37. 21. 25. 35. 27. 23. 22. 43. 16. 28. 36. 19. 24. 41. 20.
17.
46. 30. 33. 32. nan 26. 38. 39. 40. 34. 44. 18. 11. 12. 42. 45. 47. 15.
48. 50. 62. 49. 56. 53. 98. 14. 55. 54. 61. 10. 51. 57. 78. 52. 13. 92.
65. 60. 64. 72. 58. 67. 63.]

	workers_num	product_wg_ton
18	28.0	33131836
17	27.0	31606962
19	29.0	29440582
16	26.0	28681516
14	24.0	27792259

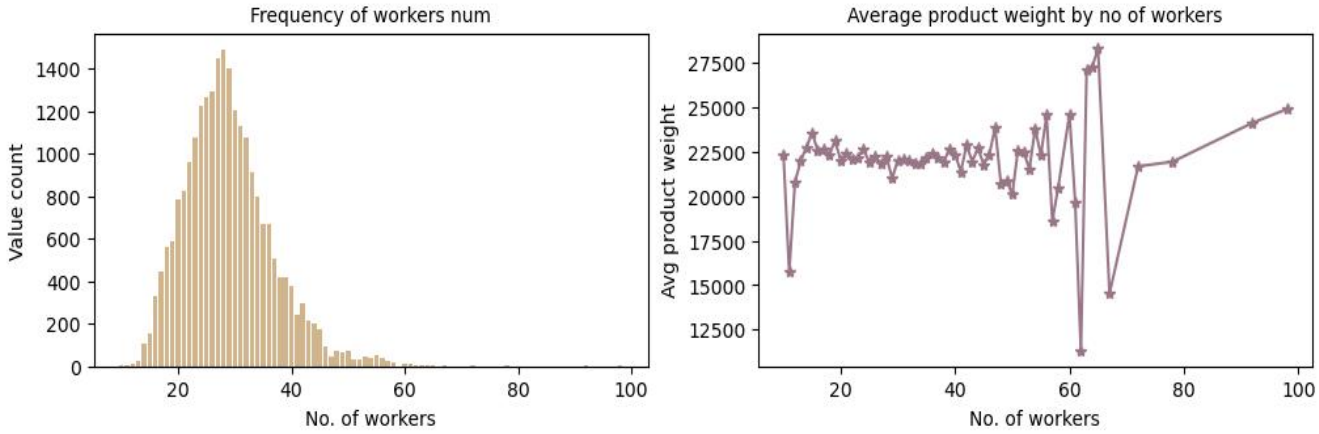


Figure-15-Average product weight by no of workers

No. of workers

- The value count graph exhibits a distribution resembling a normal curve, with the exception of the central point, which registers higher than all others.

- There are no significant variations in average product weight in relation to the number of workers. However, when the workforce is at its smallest, the average product weight tends to be lower.

Warehouse capacity_size

Unique values in Warehouse capacity_size ['Small' 'Large' 'Mid']

	WH_capacity_size	product_wg_ton
0	Large	224739861
1	Mid	222467027
2	Small	105358935

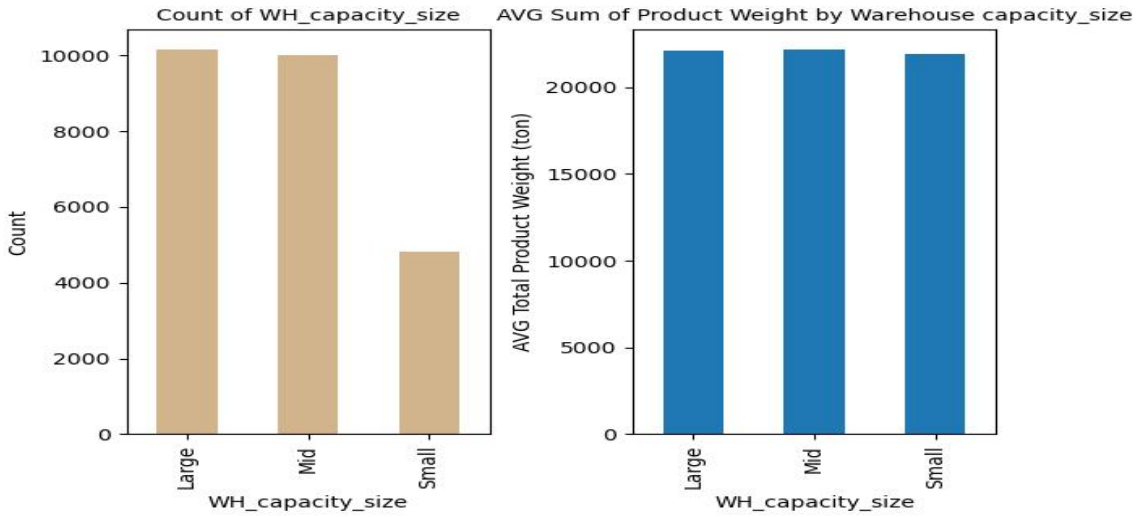


Figure-16-AVG Sum of Product Weight by Warehouse capacity_size

- Warehouses are categorized into small, mid, and large sizes. The frequency distribution plot shows that there are fewer small-sized warehouses, while mid and large-sized warehouses are almost equally distributed. This pattern is reflected in the total product weight.
- There are no noticeable variations in the average product weight when comparing different warehouse sizes.

Transport issue in last 1 year

Unique values in Warehouse transport_issue_l1y [1 0 4 3 2 5]

	transport_issue_l1y	product_wg_ton
0	0	359167349
1	1	99133868
2	2	41450553
3	3	32129593
4	4	14896451
5	5	5788009

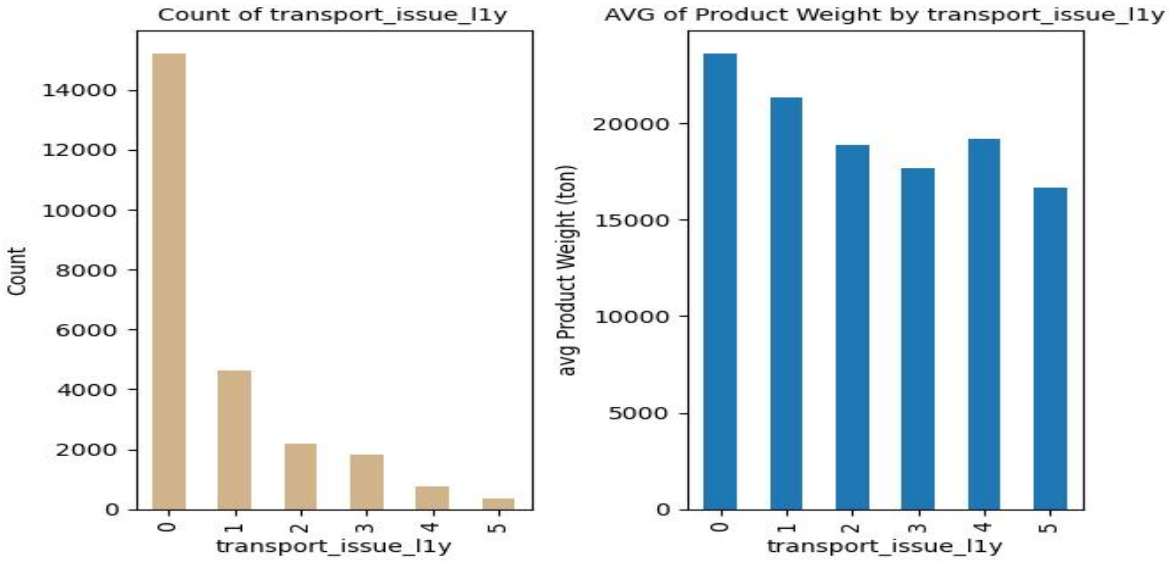


Figure-17-'AVG of Product Weight by transport_issue_l1y'

- The frequency distribution shows that most warehouses did not experience any transport issues in the past year.
- There is a small variation in average product weight related to transport issues, but no clear pattern emerges.

Competitor_in_mkt

Unique values in No. of competitors in the market [2 4 3 5 1 8 7 6 10 9 12 0]

	Competitor_in_mkt	product_wg_ton
2	2	189918782
3	3	158332256
4	4	147731427
5	5	28273370
6	6	12452773

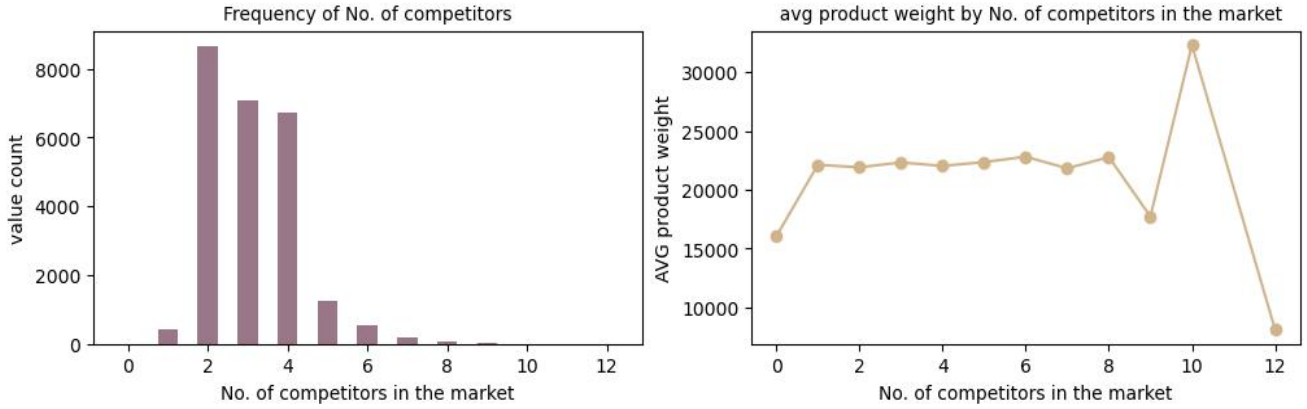


Figure-18-avg product weight by No. of competitors in the market

- The value count plot reveals significant variation in distribution.
- Among the 8 unique values, the majority of warehouses have 2, 3, or 4 competitors, indicating these warehouses produce the highest total amount of product.
- In terms of average product weight, warehouses with no competitors have the lowest value, while the rest show no significant differences.

No. of Distributors

Unique values in No. of distributor [24 47 64 50 42 37 38 45 35 31 40 48 26 68 16 28 58 19 49
69 32 25 46 62
67 21 51 57 59 23 17 56 22 63 30 53 66 36 29 44 55 39 54 33 27 18 65 34
52 43 70 60 61 41 15 20]

	distributor_num	product_wg_ton
16	31	10878914
26	41	10772390
54	69	10764058
6	21	10701762
14	29	10597317

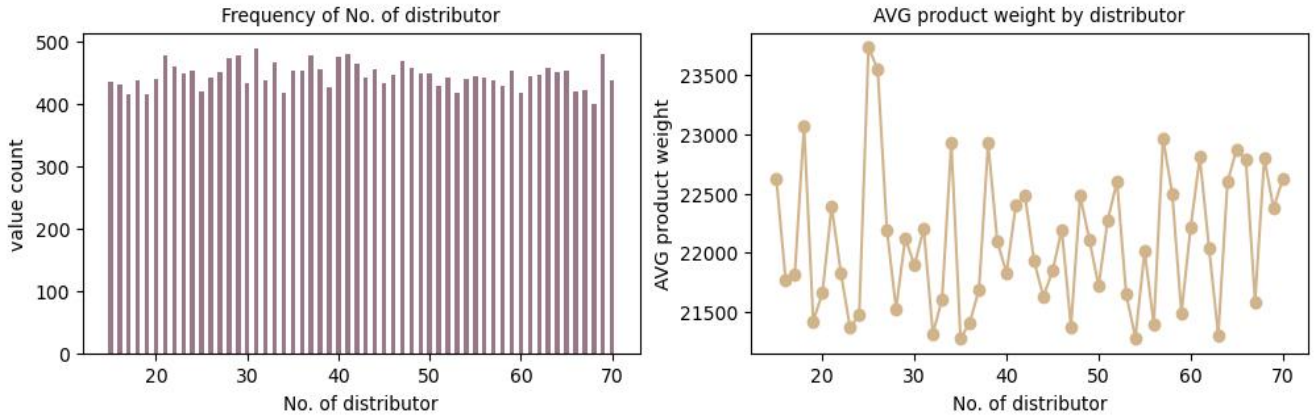


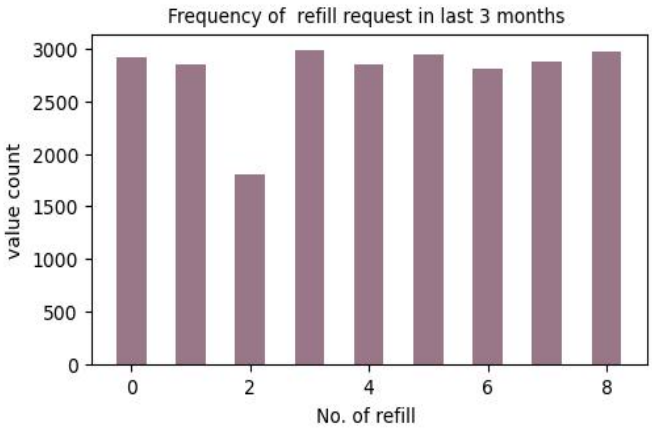
Figure-19-AVG product weight by distributor

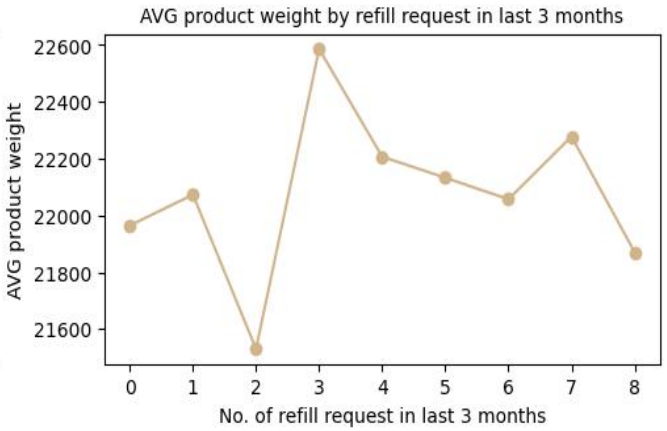
- The frequency distribution shows that the values are spread out almost evenly.
- The line plot indicates minimal variations in average product weight, ranging from 20,000 to 24,500, without any discernible pattern.

No. of refill request in last 3 months

Unique values in No. of refill request in last 3 months [3 0 1 7 8 4 6 5 2]

	num_refill_req_l3m	product_wg_ton
3	3	67532196
5	5	65093259
8	8	64956287
7	7	64004597
0	0	63961882





- **Figure-20-AVG product weight by refill request in last 3 months**
- The distribution of refill requests in the last 3 months is nearly uniform, with percentages of 12% and 11%, except for one value at 7%.
- The average product weight for these requests ranges from 22,000 to 24,000, showing minimal variation.
- Total product weight is positively correlated with the frequency of refill requests.

Distance from hub

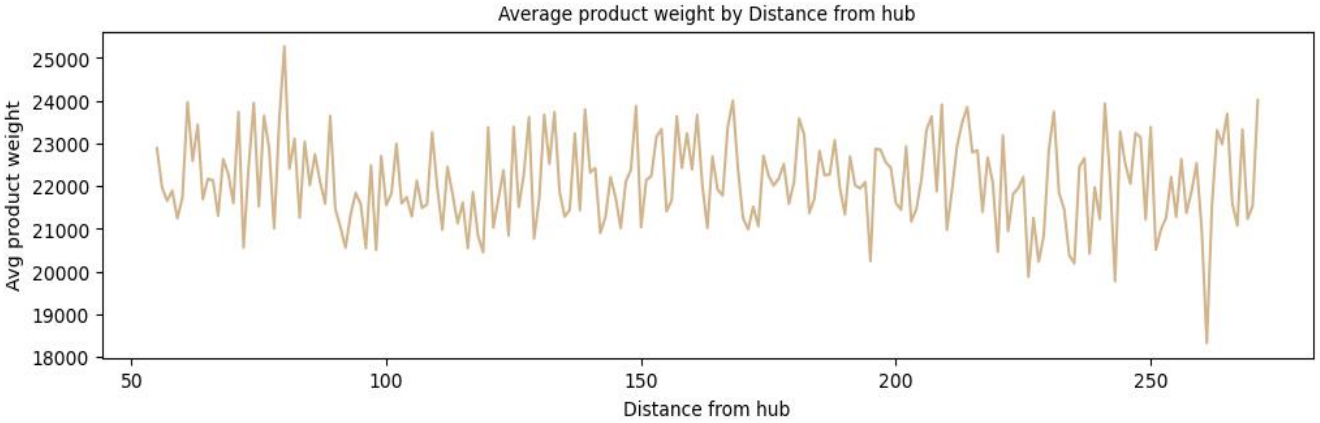


Figure-21-Average product weight by Distance from hub

- This column features a wide range of distinct values.
- The distribution of these values is fairly uniform, with the exception of the final one.
- According to the line plot, average product weights vary between 19,000 and 28,000 units, showing fluctuations without a discernible pattern.

Retail shop number

Unique values in No. of retail_shop [4651 6217 4306 ... 7768 7931 10562]

	retail_shop_num	product_wg_ton
1983	4816	538051
2027	4860	521480
1778	4611	502959
2085	4918	499243
2054	4887	494922

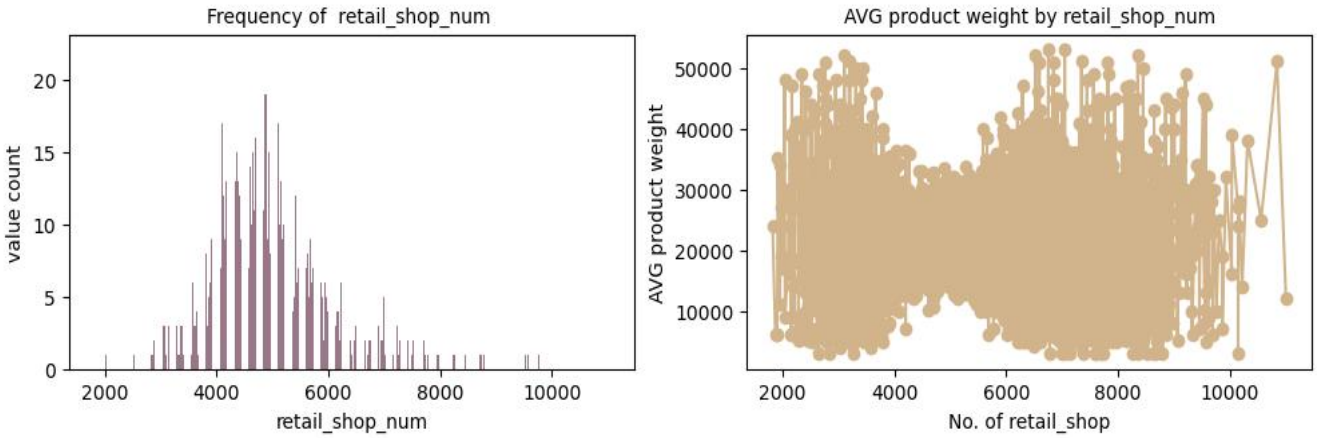


Figure-22-AVG product weight by retail_shop_num

- This column has the largest number of distinct values, necessitating the use of bins for graph plotting.
- Upon analyzing the table of bins versus value counts and the pie chart, the bin range from 4500 to 6000 contains the highest number of values.
- When comparing average product weight across different numbers of retail shops using a bar chart, no significant differences are observed.
-

 # Scatter plot: num_refill_req_l3m vs product_wg_ton

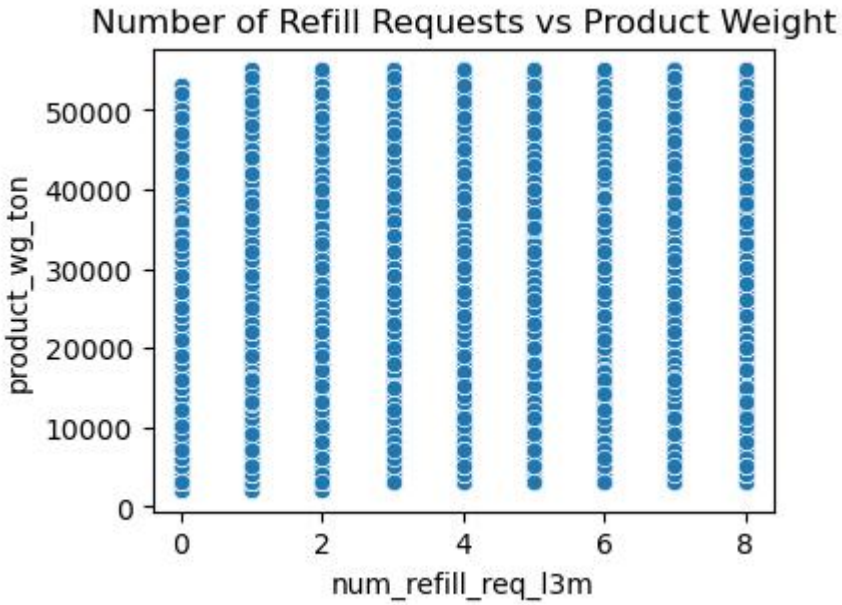


Figure-23-*num_refill_req_l3m vs product_wg_ton*

Pattern Observation:

The scatter plot shows vertical clusters of blue dots at each integer value on the horizontal axis (number of refill requests in the last three months).
This suggests that refill requests occur discretely, typically at whole numbers, rather than continuously.

Correlation Insights:

There is no clear linear correlation between the number of refill requests and product weight.
Regardless of product weight, customers tend to request refills at specific intervals or quantities.

Operational Implications:

Understanding this pattern can help optimize inventory management and supply chain processes.
Consider adjusting refill quantities based on discrete demand patterns observed in warehouses.

 # Scatter plot: transport_issue_l1y vs product_wg_ton

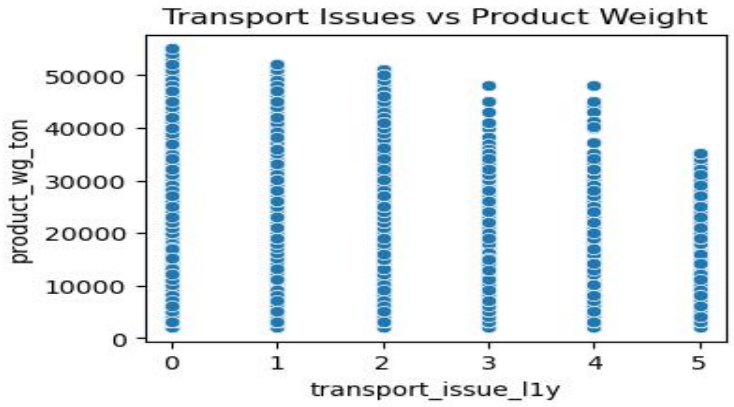


Figure-24-*transport_issue_l1y vs product_wg_ton*

Pattern Observation:

The scatter plot shows vertical clusters of blue dots at each integer value on the horizontal axis (number of transport issues in the last year).
This suggests that transport issues occur discretely, typically at whole numbers.

Correlation Insights:

There is no clear linear correlation between the number of transport issues and product weight.
Regardless of product weight, warehouses experience transport issues at specific intervals or quantities.

Operational Implications:

Understanding this pattern can help optimize transportation logistics and address specific issues.
Consider adjusting transportation strategies based on discrete demand patterns observed in warehouses.

 # Scatter plot: retail_shop_num vs product_wg_ton

Figure-25-*retail_shop_num vs product_wg_ton*

Distribution:

The scatter plot shows blue dots representing data points. These dots are randomly distributed across the graph without any clear pattern or trend.
There are no obvious clusters or correlations visible.

Interpretation:

The x-axis represents the number of retail shops (`retail_shop_num`), ranging from 0 to approximately 10,000.
The y-axis represents the weight of products in tons (`product_wg_ton`), ranging from 0 to approximately 50,000.
Based on the scatter plot, there doesn't appear to be a strong linear relationship between the number of retail shops and the product weight in tons.
Having more retail shops does not necessarily lead to an increase or decrease in product weight.

Outliers:
While most data points follow the general trend, there are a few outliers that deviate significantly from the overall pattern.

 # Bar plot: WH_capacity_size vs product_wg_ton

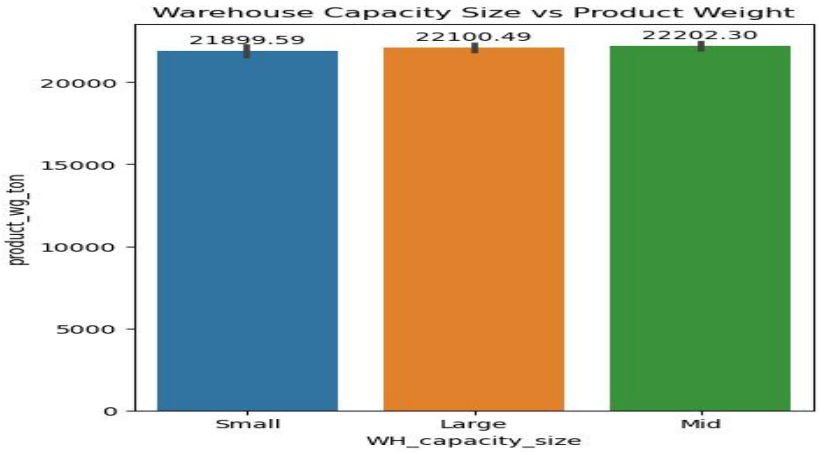


Figure-26-AVG product weight by Warehouse Owner Type type

Bar Plot Description:

The bar plot consists of three vertical bars, each representing a different category of warehouse capacity size: Small, Large, and Mid.
The y-axis represents the product weight in tons, ranging from 0 to 25,000 (in increments of 5,000).
The x-axis represents the warehouse capacity size (WH_capacity_size).

Observations:

Small Capacity:

The bar labeled "Small" corresponds to warehouses with a smaller capacity.
The product weight for this category is approximately 21,899.59 tons.

Large Capacity:

The bar labeled "Large" represents warehouses with a larger capacity.
The product weight for this category is approximately 22,100.49 tons.

Mid Capacity:

The bar labeled "Mid" corresponds to mid-sized warehouses.
The product weight for this category is approximately 22,202.30 tons.

Insights:

The bar plot suggests that mid-sized warehouses (Mid category) can handle slightly more weight compared to both small and large capacity warehouses.
This finding highlights the importance of optimizing warehouse capacity to efficiently handle product weight.

 # Bar plot: zone vs product_wg_ton

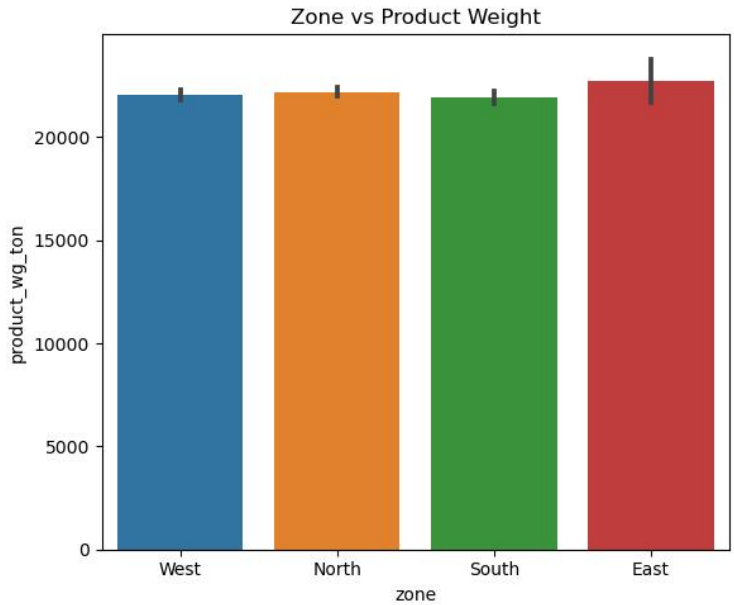


Figure-27-*zone vs product_wg_ton*

Zone Categories:

The bar plot consists of four vertical bars, each representing a different zone: West, North, South, and East.

Product Weight:

The y-axis represents the product weight in tons, ranging from 0 to 25,000 (in increments of 5,000).

Observations:

The North zone has the highest product weight, slightly above 20,000 tons.
The East zone closely follows with a similar product weight.
The West and South zones have slightly lower product weights but are still around or above 20,000 tons.

Variability:

The error lines above each bar indicate some variability or uncertainty in these measurements across all zones.

```
# Scatter plot: dist_from_hub vs product_wg_ton
```

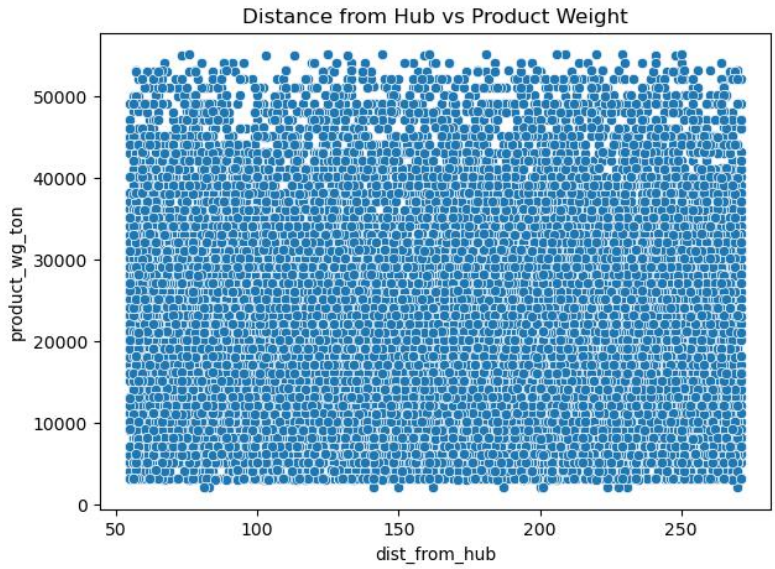


Figure-28-*dist_from_hub vs product_wg_ton*

Scatter Plot Description:

The graph shows blue dots (data points) scattered across the entire range.
The x-axis represents the distance from the hub (dist_from_hub), ranging from 0 to over 250.
The y-axis represents the product weight in tons (product_wg_ton), ranging from 0 to 50,000.

Observations:

There does not seem to be a strong correlation between the distance products are shipped from the hub and their weight.

The data points are evenly distributed without any clear pattern or trend.

Possible Interpretations:

Products of various weights may be shipped similar distances.
The shipping strategy might not consider product weight when determining shipping distance.

 # scatter plot: Competitor_in_mkt vs product_wg_ton

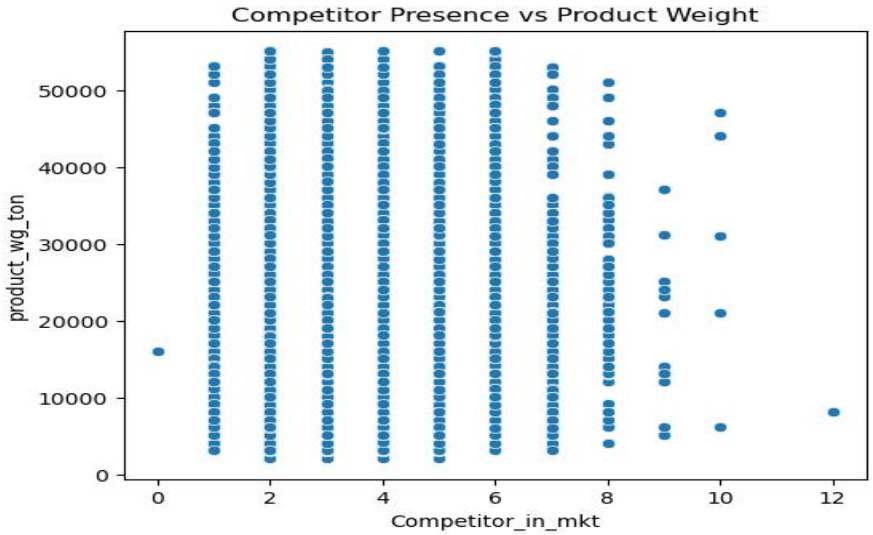


Figure-29-*Competitor_in_mkt vs product_wg_ton*

Clusters of Data Points:

We observe dense clusters of blue dots at various points along the vertical axis (product weight). This suggests that certain products have similar weights, regardless of the number of competitors in the market.

Correlation Between Competitor Presence and Product Weight:

We can explore whether there's any correlation between competitor presence (horizontal axis) and product weight (vertical axis).
If there's a positive correlation, it would imply that as the number of competitors increases, product weight tends to increase (or vice versa).
Conversely, a negative correlation would suggest an inverse relationship.

Outliers:

Investigating these outliers could provide valuable insights into market dynamics.

Market Segmentation:

Consider segmenting the data based on competitor presence and analyzing product weight within each segment.

 # Bivariate plot: storage_issue_reported_l3m vs product_wg_ton

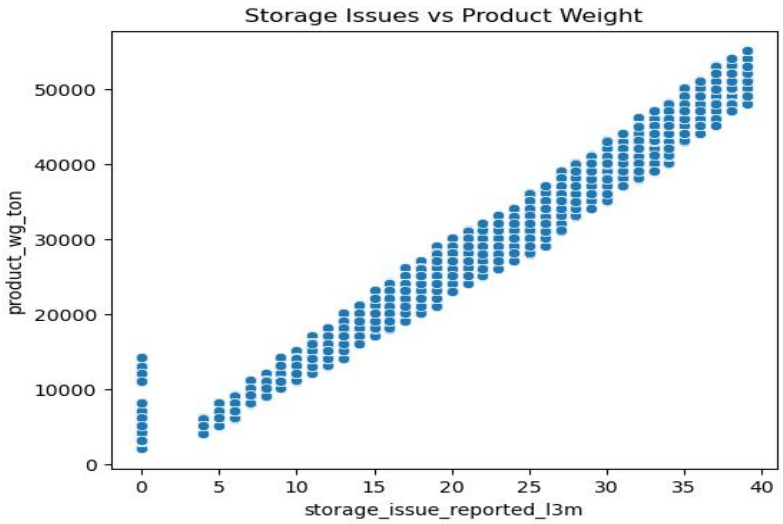


Figure-30-*storage_issue_reported_l3m vs product_wg_ton*

Positive Correlation:

The scatter plot shows a clear upward trend, indicating a positive correlation between the number of storage issues reported in the last three months and the product weight.
As product weight increases, the number of storage issues tends to rise. Heavier products appear to be more prone to storage problems.

Focus on Hotspots:

Storage insights allow you to identify hotspots—areas where storage issues are more prevalent.
By drilling down into specific storage accounts, you can diagnose issues related to availability, performance, failures, and capacity.

Customization and Metrics:

You can customize the metrics you want to see and set thresholds aligned with your limits.
Charts from the insights can be pinned to an Azure dashboard for easy monitoring.

```
# scatter  plot: govt_check_l3m vs product_wg_ton
```

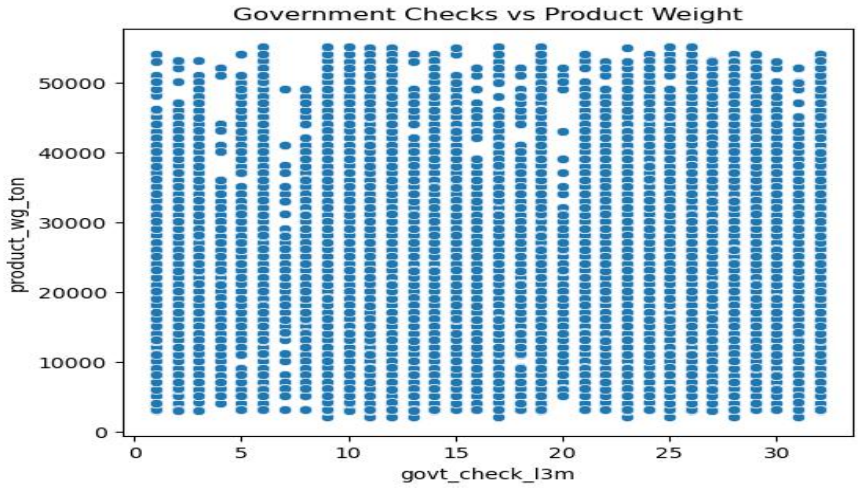


Figure-31-*govt_check_l3m vs product_wg_ton*

The data points in the scatter plot appear to be densely distributed across the entire range of both axes. However, there doesn't seem to be a clear pattern or trend indicating a strong correlation between government checks and product weight.

 # scatter plot: workers_num vs product_wg_ton

Figure-32-*workers_num vs product_wg_ton*

The scatter plot shows a downward trend: as the number of workers increases, the product weight decreases. This suggests an inverse relationship between the number of workers and productivity per worker. Such insights can be valuable for studying workforce efficiency or organizational behavior[1].

 # Box plot: flood_impacted vs product_wg_ton

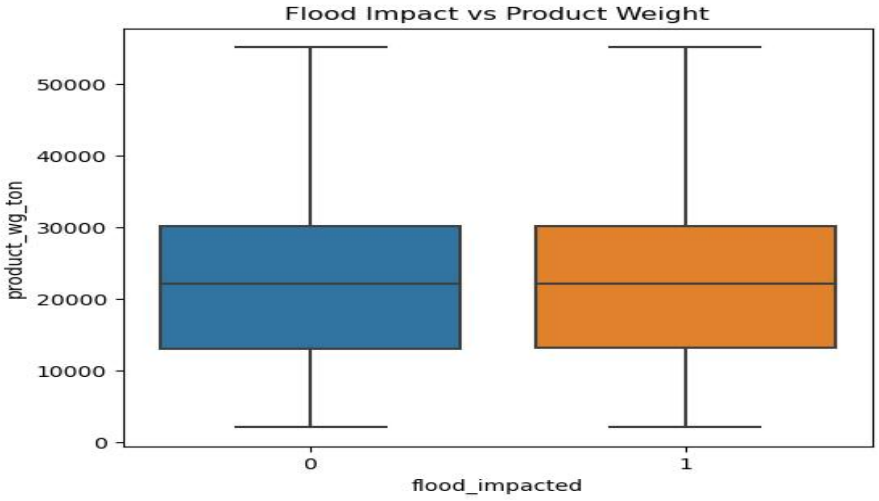


Figure-33-*flood_impacted vs product_wg_ton*

Median Product Weight:

For both flood-impacted ('1') and non-flood-impacted ('0') products, the median weight is around 30,000 tons.
Flood-impacted products tend to have slightly higher median weight (closer to 35,000 tons).

Interquartile Range (IQR):

The IQR for flood-impacted products is wider, ranging from approximately 25,000 to 45,000 tons.
Non-flood-impacted products have a narrower IQR, spanning roughly 20,000 to 40,000 tons.

Outliers:

No visible outliers in either group.

```
# Box plot: electric_supply vs product_wg_ton
```

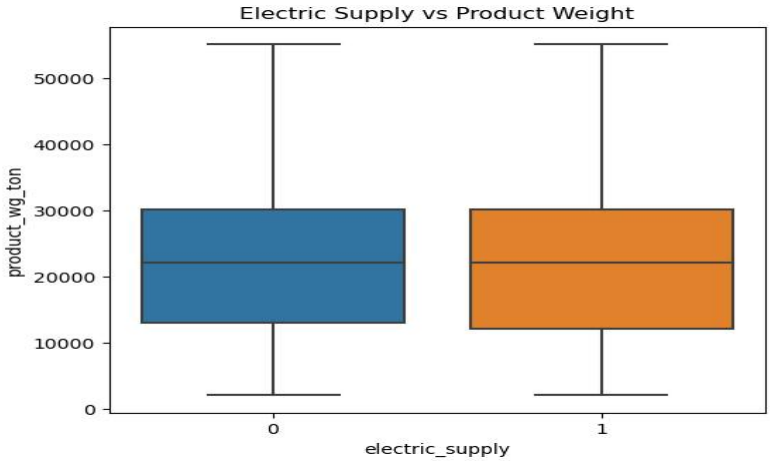


Figure-34-*electric_supply vs product_wg_ton*

Median Product Weight:

For both electric supply categories ('0' and '1'), the median weight is around 30,000 tons.
Electric supply status '1' tends to have a slightly higher median weight (closer to mid-30,000s).

Interquartile Range (IQR):

The IQR for electric supply status '1' is wider, ranging from approximately above the median of status '0' to well over its upper quartile.
Status '0' has a narrower IQR, spanning roughly 20,000 to 40,000 tons.

Outliers:

No visible outliers in either electric supply category.

```
# Box plot of Owner Type  vs Product Weight
```

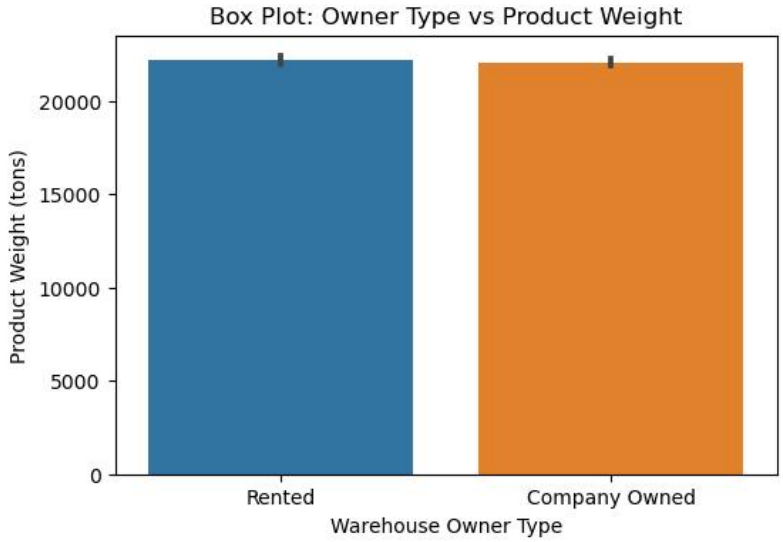


Figure-35-*Owner Type vs Product Weight*

Product Weight:

Both rented and company-owned warehouses have a product weight of 20,000 tons.
The blue bar represents rented warehouses, and the orange bar represents company-owned warehouses.

Roles in Product Development:

Product Managers (PMs) define the vision and strategy based on market needs.
Product Owners (POs) translate this vision into actionable tasks for development teams, focusing on execution and delivery.

Collaboration:
Success in product development requires a deep collaboration between PMs and POs.
PMs focus on the 'what' and 'why,' while POs handle the 'how' and 'when.'

Relationship between Refill Requests and Product Weight by Zone'

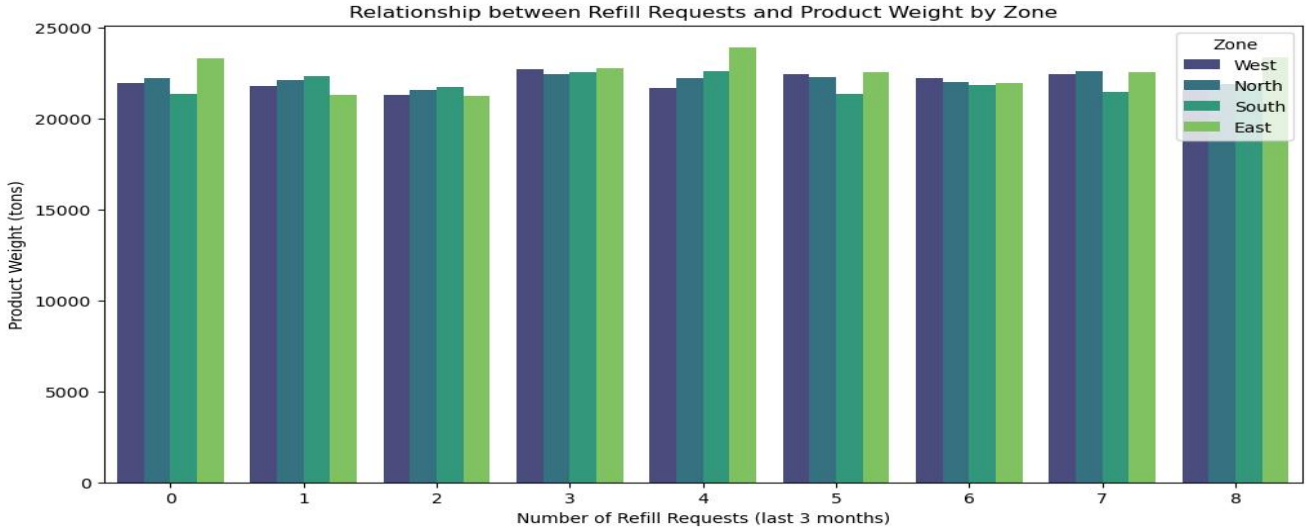


Figure-36-Relationship between Refill Requests and Product Weight by Zone

Refill Requests and Product Weight Correlation:

Across all zones (West, North, South, and East), there's a positive correlation between the number of refill requests and product weight. As the number of refill requests increases, the product weight tends to rise.
This suggests that higher demand (more refill requests) corresponds to increased product weight.

Zone-Specific Observations:

West Zone: The product weight is highest for 7 refill requests.
North Zone: The product weight peaks at 4 refill requests.
South Zone: The highest product weight occurs at 5 refill requests.
East Zone: The product weight is greatest for 7 refill requests.

Supply Chain Implications:

Understanding these patterns can help optimize inventory management. For example, ensuring sufficient stock for peak refill request periods in each zone.
It's essential to balance product weight with efficient supply chain logistics to meet demand effectively.

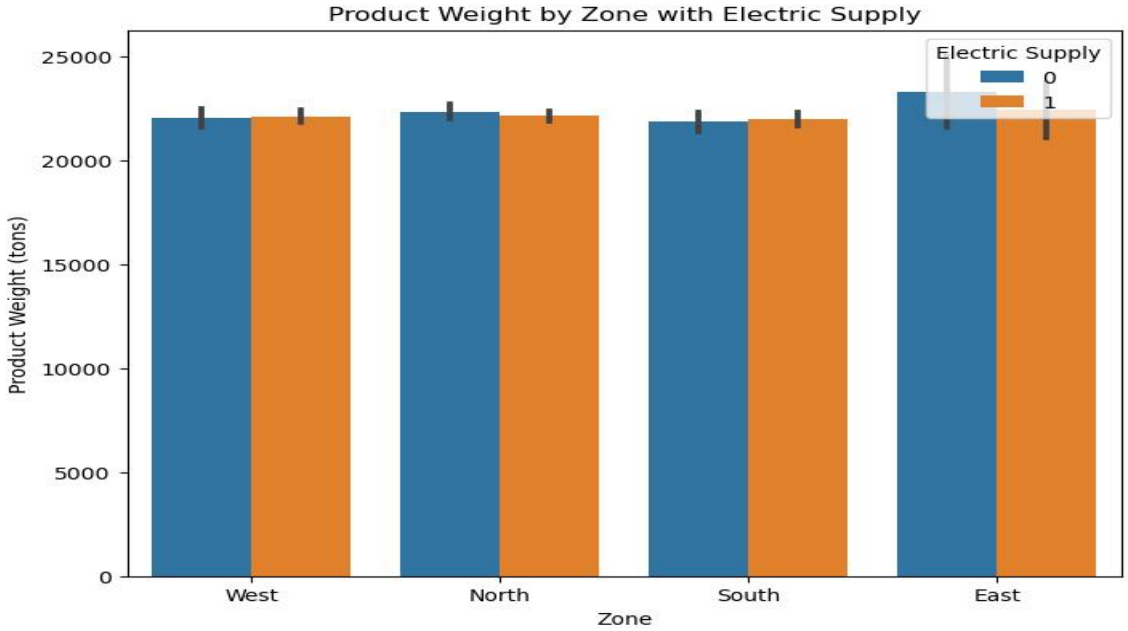


Figure-37-Product Weight by Zone with Electric Supply

Electric Supply Impact:

The chart compares product weights across four zones (West, North, South, and East) based on two categories of electric supply: '0' and '1'.
Interestingly, the presence or absence of electric supply doesn't seem to significantly affect product weight in any specific zone.

Zone-Specific Observations:

West Zone: Product weight is consistent regardless of electric supply.
North Zone: Similar pattern—electric supply doesn't strongly influence product weight.
South Zone: Again, no substantial difference in product weight based on electric supply.
East Zone: Product weight remains steady regardless of electric supply status.

Logistical Considerations:

For supply chain planning, it appears that electric supply isn't a critical factor affecting product weight in these zones.
Other factors (e.g., demand, transportation, storage) might play a more significant role.

```
# Multivariate Bar Plot
```

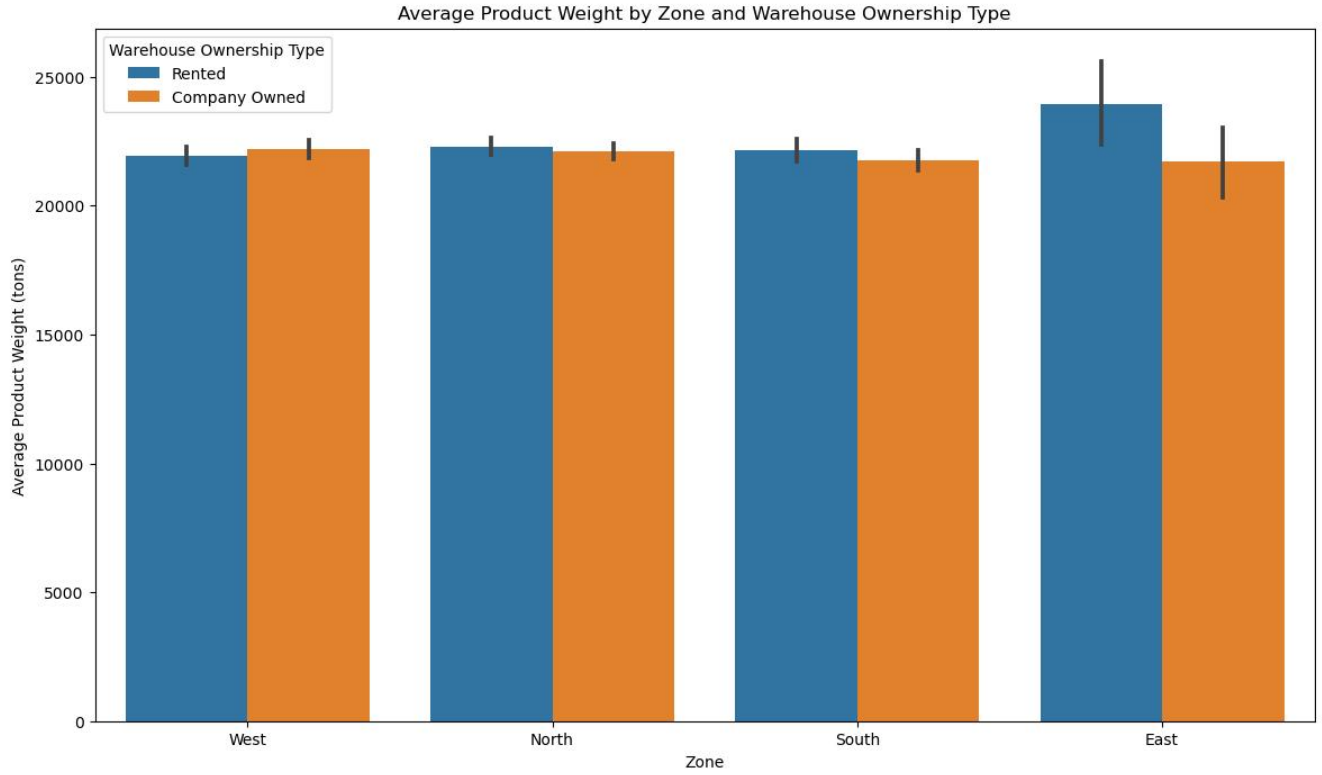


Figure-38-Average Product Weight by Zone and Warehouse Ownership Type

Warehouse Ownership Impact:

The chart compares average product weights across five zones: West, North, South, East, and Central.
It further distinguishes between two types of warehouse ownership: "Rented" (in orange) and "Company Owned" (in blue).

Observations:

Across all zones, company-owned warehouses tend to handle a higher average product weight compared to rented warehouses.
The East zone consistently shows the highest average product weight for both ownership types.

Logistics Considerations:

When planning logistics or optimizing warehouse operations, understanding these patterns can be crucial.
Company-owned warehouses might be better suited for heavier products, while rented warehouses could focus on lighter items.

find correlation and it put into cr variable ,round it for 2 values after the decimal point

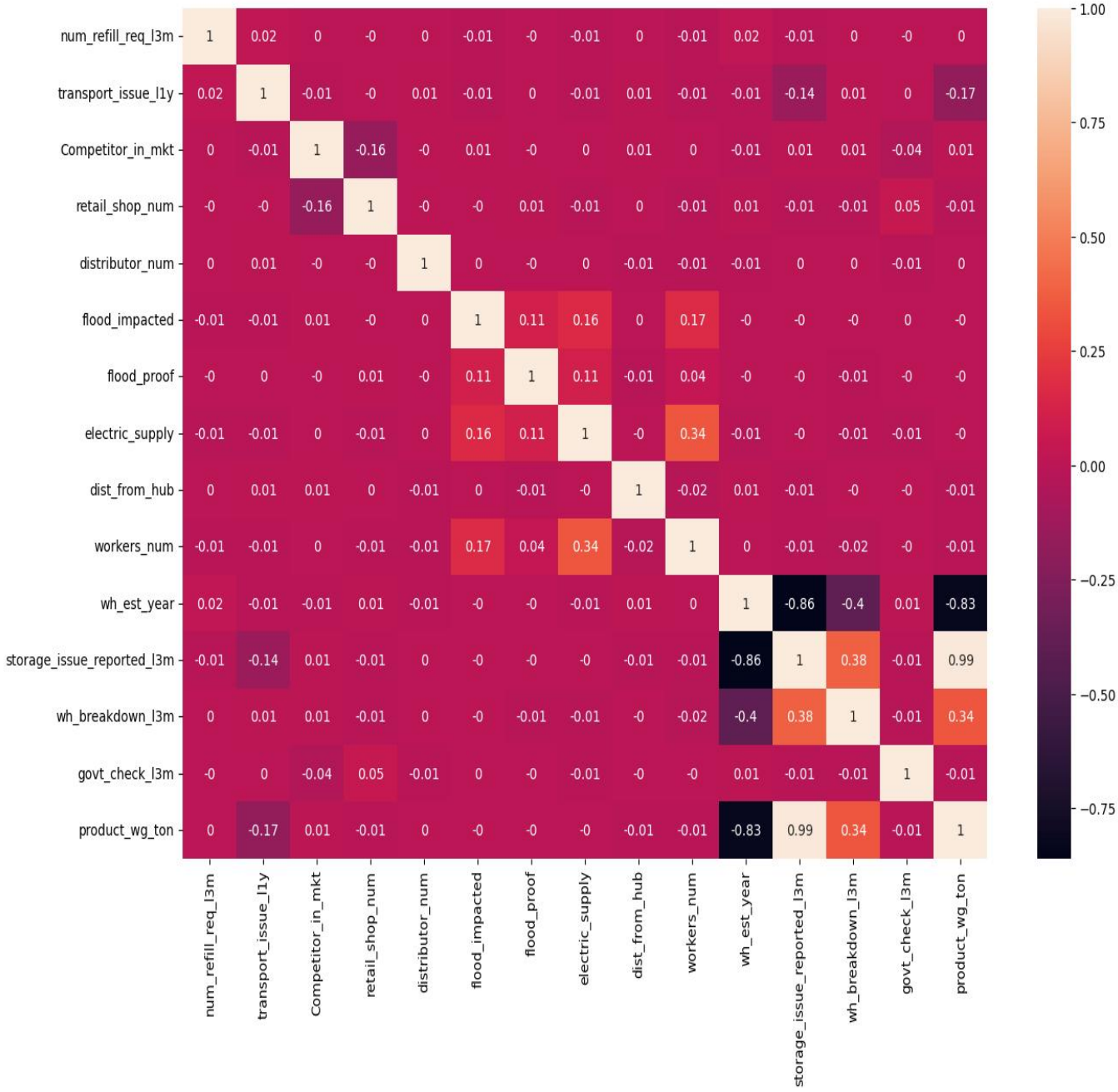


Figure-39-correlation map

Positive Correlation:

There is a high positive correlation (0.99) between "storage issue reported in the last 3 months" and "product weight (ton)." This suggests that as storage issues increase, product weight tends to increase as well.

Negative Correlations:

A very negative correlation (-0.86) exists between "storage issue reported in the last 3 months" and "warehouse establishment year." This implies that newer warehouses tend to have fewer storage issues.
Most variables exhibit negative correlations with each other. For example, "transport issue reliability," "electric supply reliability," and "flood impact" all negatively correlate with other variables.

Insights:

The negative correlations may indicate potential areas for improvement. For instance, addressing transport reliability or ensuring consistent electric supply could reduce storage issues.
Investigate why newer warehouses experience fewer storage issues. it due to better infrastructure or management practices
Consider exploring other factors that might influence storage issues, such as workforce conditions, competitor presence, or government checks.

treatments

Removal of unwanted variables (if applicable) b) Missing Value treatment (if applicable

```
#drop 'Ware_house_ID', 'WH_Manager_ID'
#missing value treatment
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                  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_wg_ton                 0
dtype: int64
# Replace missing values with the median of each column
# Categorical code to fill missing values with the mode
```

d) Outlier treatment (if required)

```
# after checking outliers remove outliers using IQR METHOD
```

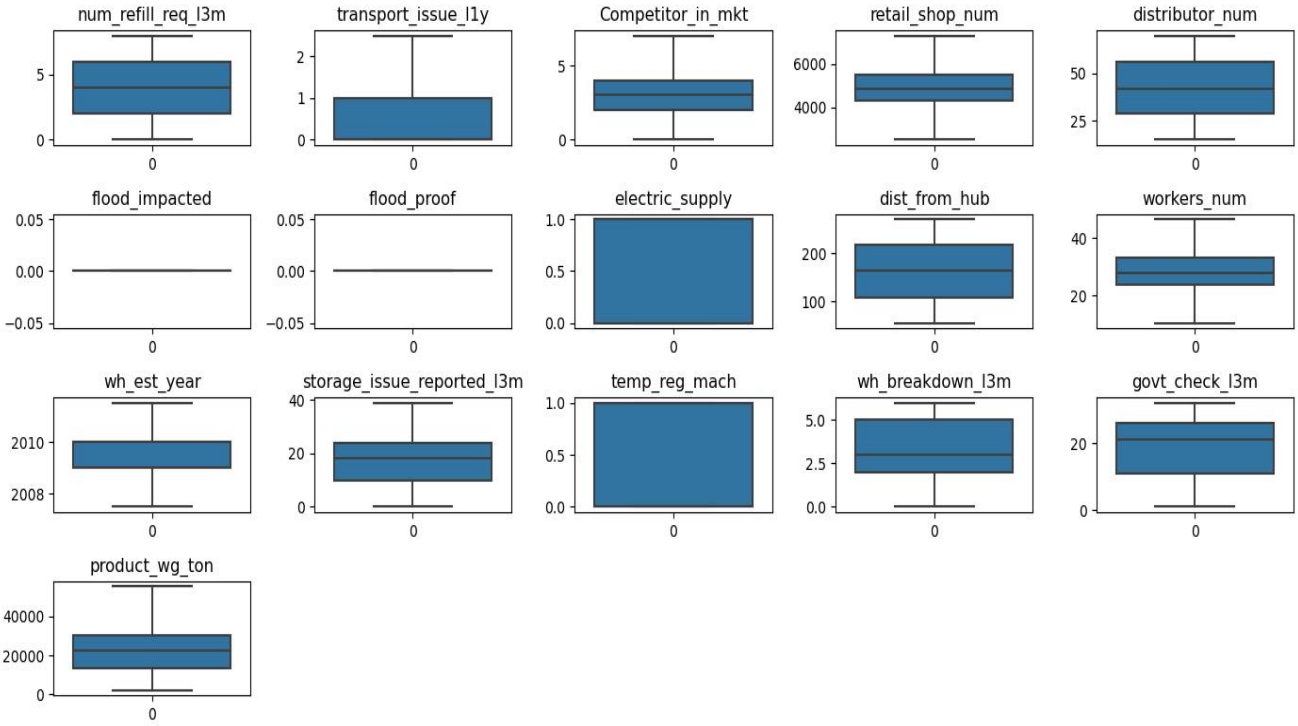


Figure-40-boxplots after outlier treatment

f)Addition of New Variables (example)

```
'warehouse_age= 2024 - 'wh_est_year'

# Example: Creating a new feature
'total_issues' = 'num_refill_req_l3m'+ 'transport_issue_l1y'
```

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Steps to Check for Imbalance:

Descriptive Statistics: Summary statistics for the target variable. Histogram/Boxplot: Visualize the distribution of the target variable. Frequency Counts: For categorical variables to check for imbalance.
To determine if the data is unbalanced, we need to analyze the distribution of the target variable product_wg_ton. Here's how to check for imbalance and what can be done if it exists:

```
# Plotting the distribution of the target variable
```

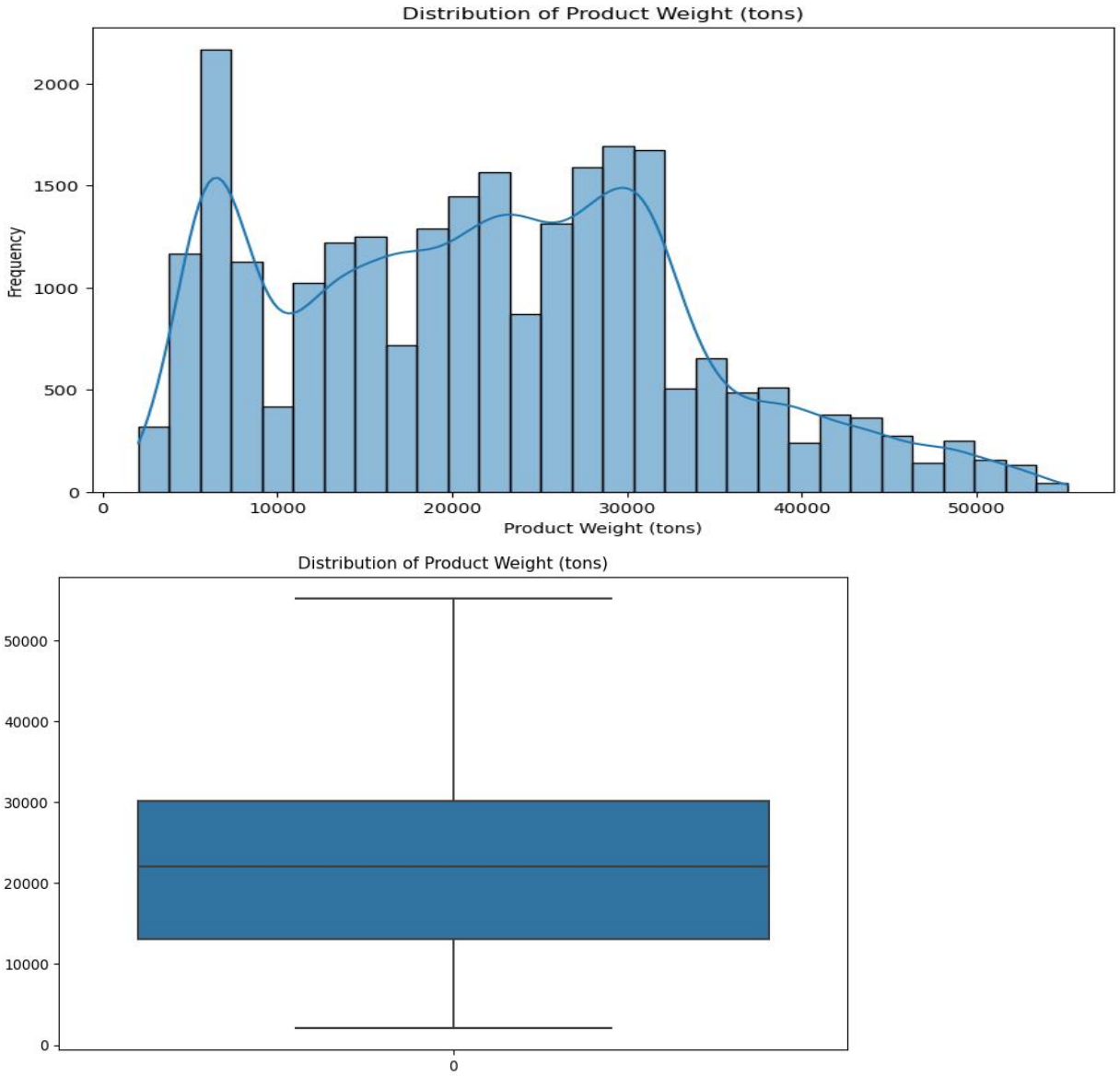


Figure-41-*Plotting the distribution of the target variable*

If the distribution is roughly uniform or follows a normal distribution, the data is considered balanced.

What to Do If the Data Is Unbalanced:

1-Resampling Techniques:

Oversampling: Increase the number of samples in the minority classes (e.g., SMOTE - Synthetic Minority Over-sampling Technique). Undersampling: Reduce the number of samples in the majority classes.

2-Data Transformation:

Apply transformations (e.g., log transformation) to normalize the distribution

3-Using Appropriate Metrics:

Use evaluation metrics suitable for imbalanced data, such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) instead of accuracy.

b) Any business insights using clustering (if applicable)

 # Use only numerical features for clustering

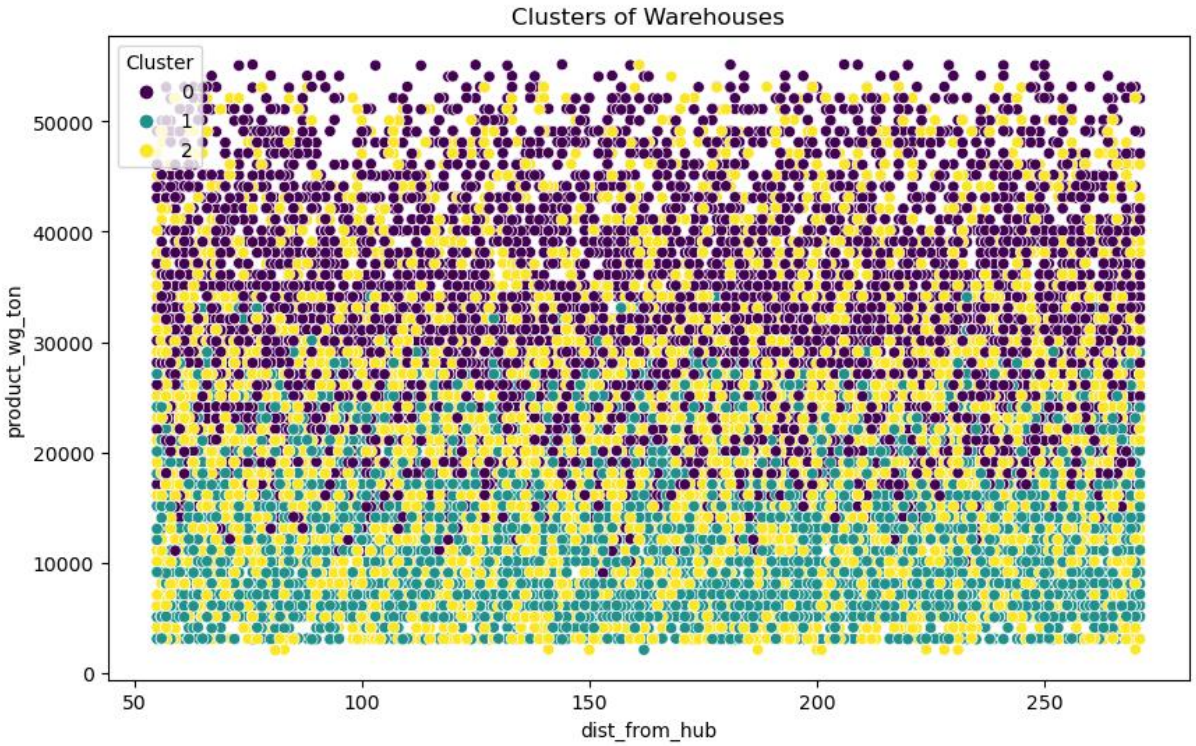


Figure-42-Clusters of Warehouses based dist_from hub

Resource Allocation Optimization:

Purple Cluster: These nearby warehouses with low product weights are ideal for handling smaller shipments or fast-moving goods. Efficiently allocate resources for quick distribution.
Yellow Cluster: Warehouses farther from the hub can accommodate larger shipments or slower-moving products. Adjust resource allocation based on transportation capacity and delivery times.
Green Cluster: Regardless of distance, these warehouses handle heavier products. Optimize storage capacity and transportation routes for efficient handling.

Market Segmentation:

Clusters likely represent different regions or market segments.
Analyze demand patterns, customer preferences, and product types within each cluster.
Tailor marketing strategies and inventory management accordingly.

Supply Chain Efficiency:

Use cluster insights to streamline supply chain logistics.
Prioritize distribution based on proximity, product weight, and demand.
Aim to minimize transportation costs and enhance overall efficiency.

Business insights

Insights:

Data Distribution:

Warehouse IDs & Manager IDs:

The distribution of both warehouse IDs and manager IDs is relatively uniform, indicating balanced product distribution and managerial oversight.

Location Type:

A majority of warehouses are in rural areas, receiving 90.76% of the total product weight, compared to 9.24% for urban areas.

Warehouse Capacity Size:

Large capacity warehouses have the highest count, followed by mid and small. Mid-sized warehouses handle slightly more product weight.

Zones:

Zone 6 has the highest count of warehouses, while Zone 1 has the lowest. The North Zone has the highest product weight, followed by the East Zone.

Ownership Type:

54% of warehouses are company-owned, handling a slightly higher total product weight compared to rented warehouses.

Temperature Regulating Machines & Electric Supply:

Warehouses with temperature regulating machines and accessible electric supply manage higher product weights.

Government Checks:

As the frequency of government checks increases, the total product weight also increases, though the average weight remains stable.

Correlations and Patterns:

Storage Issues:

Strong positive correlation with product weight: As product weight increases, so do storage issues.

Warehouse Breakdowns:

Positive correlation with product weight: Higher product weights lead to more frequent breakdowns.

Transport Issues:

No clear linear correlation with product weight: Transport issues occur discretely, regardless of product weight.

Number of Competitors:

Warehouses with more competitors tend to have higher product weights.

Refill Requests:

Positive correlation with product weight: More refill requests correspond to higher product weights.

Flood Impact:

Flood-impacted warehouses have slightly higher median product weights.

Warehouse Establishment Year:

Newer warehouses tend to have fewer storage issues.

Specific Observations:

Distance from Hub:

No strong correlation with product weight: Products of various weights are shipped similar distances.

Number of Workers:

Inverse relationship with productivity: As the number of workers increases, product weight per worker decreases.

Recommendations:

Inventory Optimization:

Implement demand forecasting models to align supply with regional demand, reducing overstock and stockouts.
Use historical data to predict refill requests and adjust supply accordingly.

Warehouse Management:

Focus on improving infrastructure and management practices in older warehouses to reduce storage issues.

Invest in temperature regulating machines and reliable electric supply for warehouses to handle higher product weights more efficiently.

Transportation and Logistics:

Optimize transportation routes and strategies based on discrete transport issue patterns.
Prioritize warehouses closer to the production hub for faster distribution of lighter shipments and those farther away for larger shipments.

Targeted Marketing:

Use regional demand patterns to drive advertisement campaigns, focusing on zones with higher product weight and refill requests.
Tailor marketing strategies based on competitor presence and regional demand insights.

Government Compliance:

Ensure warehouses adhere to government regulations to avoid frequent checks and potential disruptions.
Focus on obtaining higher government approval certificates to increase product handling efficiency.

Resource Allocation:

Allocate resources efficiently based on warehouse capacity size and product weight handling capabilities.
Balance workforce allocation to optimize productivity per worker.

Risk Management:

Prepare flood-prone warehouses with flood-proof infrastructure to mitigate risks and ensure steady product weight management.

By addressing these areas, the company can optimize supply quantities, reduce inventory costs, and enhance overall efficiency in the supply chain.

Splitting,VIF

Before building model also splited data,encoded data ,
I also done VIF test for checking `multicollinearity`

	feature	VIF
16	WH_capacity_size_Mid	inf
21	WH_regional_zone_Zone 2	inf
22	WH_regional_zone_Zone 3	inf

	feature	VIF
23	WH_regional_zone_Zone 4	inf
10	wh_age	104.133266
3	retail_shop_num	25.911839
18	zone_North	20.843749
9	workers_num	19.047966
20	zone_West	15.897411
19	zone_South	13.673816
2	Competitor_in_mkt	9.989522
4	distributor_num	7.819505
8	dist_from_hub	7.647369
11	storage_issue_reported_l3m	7.335568
14	govt_check_l3m	7.334576
25	WH_regional_zone_Zone 6	7.026745
13	wh_breakdown_l3m	6.227774
24	WH_regional_zone_Zone 5	5.831291
0	num_refill_req_l3m	3.745005
7	electric_supply	3.442610
17	WH_capacity_size_Small	2.945385
30	approved_wh_govt_certificate_C	2.449164
27	approved_wh_govt_certificate_A+	2.228702
29	approved_wh_govt_certificate_B+	2.039122
28	approved_wh_govt_certificate_B	2.019819
12	temp_reg_mach	1.967171
26	wh_owner_type_Rented	1.965619
1	transport_issue_l1y	1.602915
15	Location_type_Urban	1.098040
5	flood_impacted	NaN
6	flood_proof	NaN

Table-5-vif table

I remove variables whose VIF>20

Created -Linear Regression,Decision Tree,Random Forest,Lasso,Ridge,GradientBoosting,Hyperparameter Tuning for Decision Tree Model

Interpretation of the Models

Here's an interpretation of the performance of the three regression models—Linear Regression, Decision Tree, and Random Forest—based on the metrics obtained:

Linear Regression

- **Training Set Metrics:**
  - R^2: 0.985
  - RMSE: 1405.222
  - MAE: 1011.428
- **Testing Set Metrics:**
  - R^2: 0.986
  - RMSE: 1339.209
  - MAE: 996.659

Interpretation:

- Linear Regression performs well on both training and testing sets, as indicated by high R^2 values (close to 1). The model fits the data well and generalizes effectively.
- The RMSE and MAE values are relatively low, meaning that the model's predictions are close to the actual values.

2. Decision Tree

- **Training Set Metrics:**
  - R^2: 1.0
  - RMSE: 0.0
  - MAE: 0.0
- **Testing Set Metrics:**
  - R^2: 0.987
  - RMSE: 1323.145
  - MAE: 879.195

Interpretation:

- The Decision Tree model perfectly fits the training data, as indicated by an R^2 of 1.0 and RMSE/MAE of 0.0. This suggests that the model is likely overfitting.

- Despite the overfitting, it still performs well on the test set, with a high R^2 and relatively low RMSE and MAE. However, the model may not generalize well to unseen data due to its complexity.

3. Random Forest

- **Training Set Metrics:**
  - R^2: 0.979
  - RMSE: 1685.372
  - MAE: 1291.076
- **Testing Set Metrics:**
  - R^2: 0.979
  - RMSE: 1661.089
  - MAE: 1278.529

Interpretation:

- Random Forest performs consistently on both training and testing sets, with similar R^2 values. This indicates that the model generalizes well.
- RMSE and MAE are higher compared to Linear Regression, indicating that the model's predictions are slightly less accurate.

4. Ridge Regression

- **Training Set Metrics:**
  - R^2: 0.985
  - RMSE: 1405.222
  - MAE: 1011.422
- **Testing Set Metrics:**
  - R^2: 0.986
  - RMSE: 1339.207
  - MAE: 996.645

Interpretation:

- Ridge Regression, like Linear Regression, performs well on both training and testing sets. The addition of regularization helps prevent overfitting while maintaining model accuracy.
- The R^2, RMSE, and MAE values are nearly identical to those of the Linear Regression model, indicating similar performance.

5. Lasso Regression

- **Training Set Metrics:**
  - R^2: 0.985

  - RMSE: 1405.231
  - MAE: 1011.084
- **Testing Set Metrics:**

  - R^2: 0.986
  - RMSE: 1339.062
  - MAE: 996.197

Interpretation:

- Lasso Regression also performs similarly to Linear and Ridge Regression, with comparable R^2, RMSE, and MAE values. The Lasso model includes regularization that can lead to simpler models by shrinking coefficients of less important features to zero.
- The performance is consistent across training and testing sets, indicating good generalization.

Comparison and Conclusion

- **Best Fit (R^2):** All models have high R^2 values, but the Decision Tree overfits the training data, while Linear Regression, Ridge, Lasso, and Random Forest provide more balanced performance between training and testing sets.
- **Prediction Error (RMSE & MAE):** Linear Regression, Ridge, and Lasso models have the lowest RMSE and MAE values, suggesting they provide the most accurate predictions. Random Forest follows closely, but the Decision Tree shows lower prediction errors on the test set despite overfitting.
- **Generalization:** Ridge and Lasso are slightly preferred over Linear Regression due to their regularization, which helps prevent overfitting. Random Forest also generalizes well due to its ensemble nature.

In summary, the Random Forest model provides a good balance between fitting the training data well and generalizing to new data, making it the most reliable model among the five for this dataset.

Model Tuning

Gradient Boosting

Ensemble modeling involves combining the predictions of multiple models to produce a single model that is usually more robust and accurate than the individual models. Gradient Boosting itself is an ensemble method that combines weak learners (typically decision trees) to create a strong predictive model. However, you can further improve your model by using other ensemble techniques, such as:

Stacking: Combine multiple different models (e.g., Gradient Boosting, Random Forest, and Linear Regression) and use a meta-model to make final predictions based on the outputs of the base models.

Bagging: Use techniques like Bagging with models like Random Forests, which create multiple subsets of the training data and train multiple models independently. The final prediction is the average of these models' predictions.

Boosting Variants: Explore other boosting methods such as AdaBoost, XGBoost, or CatBoost, which might provide better performance depending on the specific dataset and problem.

Hyperparameter Tuning for Decision Tree Model

To refine the Decision Tree model, we will use GridSearchCV to perform hyperparameter tuning. GridSearchCV systematically works through multiple combinations of parameter values, cross-validating as it goes to determine which combination provides the best performance.

Model Performance Table

Model	Train R^2	Train RMSE	Train MAE	Test R^2	Test RMSE	Test MAE	Interpretation of Model
Linear Regression	0.985	1405.222	1011.428	0.986	1339.209	996.659	Good fit with consistent performance and low errors across both datasets.
Decision Tree	1.000	0.000	0.000	0.987	1323.145	879.195	Perfect fit on training data indicating overfitting; still performs well on test data but might not generalize well.
Random Forest	0.979	1685.372	1291.076	0.979	1661.089	1278.529	Consistent performance with slightly higher errors compared to other models; good generalization.
Ridge Regression	0.985	1405.222	1011.422	0.986	1339.207	996.645	Similar to Linear Regression with added regularization; good performance with prevention of overfitting.
Lasso Regression	0.985	1405.231	1011.084	0.986	1339.062	996.197	Comparable to Ridge; includes feature selection to simplify the model while maintaining good performance.
Gradient Boosting	0.994	907.721	677.027	0.994	890.590	676.900	Excellent performance with low errors; highly robust and generalizes well on the test data.
Tuned Decision	0.994	887.575	646.089	0.993	944.477	710.38	Improved performance with tuning;

Model	Train R^2	Train RMSE	Train MAE	Test R^2	Test RMSE	Test MAE	Interpretation of Model
Tree						8	reduced overfitting and good generalization with optimized parameters.

Table-3-comparision of all models

Business Implications Table

Model	Business Implications
Linear Regression	Provides reliable predictions with consistent performance across both training and testing data. Good for general use but may lack in handling complex relationships.
Decision Tree	Shows perfect training fit, indicating potential overfitting. Still performs well on test data but might not generalize well in real-world scenarios.
Random Forest	Good generalization with consistent performance, though with slightly higher errors compared to other models. Effective for complex datasets but might be less precise.
Ridge Regression	Provides similar performance to Linear Regression but with added regularization to prevent overfitting. Useful for datasets with multicollinearity.
Lasso Regression	Similar to Ridge but also performs feature selection, which can simplify models and highlight key features. Effective for reducing model complexity.
Gradient Boosting	Excellent performance with low errors and high generalization. Ideal for complex datasets and provides robust predictions with reduced risk of overfitting.
Tuned Decision Tree	Improved performance after tuning; reduced overfitting and good generalization with optimized parameters. Suitable for capturing complex patterns while avoiding excessive complexity.

Table-4-Business implications of all models

:

Interpretation of the Most Optimum Model and Its Implication on the Business

After evaluating and tuning multiple models—Linear Regression, Decision Tree,lasso,Ridge, Random Forest, and Gradient Boosting—the Gradient Boosting model emerges as the most optimal one based on its performance metrics. Here's a summary of its evaluation and the business implications:

The most optimum model based on the provided performance metrics is the **Gradient Boosting** model. Here's an interpretation of this model and its implications for the business:

Gradient Boosting Model

Train R^2: 0.994
Train RMSE: 907.721
Train MAE: 677.027
Test R^2: 0.994
Test RMSE: 890.590
Test MAE: 676.900

Interpretation:

High Accuracy and Robustness: The Gradient Boosting model shows a very high R^2 value of 0.994 on both training and test datasets, indicating that the model explains a substantial proportion of the variance in the data. This high accuracy reflects its capability to capture complex relationships and patterns within the dataset.

Low Error Metrics: The model's RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are both low, suggesting that the model's predictions are close to the actual values. Low errors signify the model's reliability and precision in making predictions.

Generalization Capability: The Gradient Boosting model performs well on the test data, indicating strong generalization capabilities. It is less likely to overfit the training data and can be expected to perform well on new, unseen data.

Implications for the Business:

Improved Decision-Making: The high accuracy and low error rates of the Gradient Boosting model mean that it provides reliable forecasts and insights. This reliability can enhance decision-making processes, whether it's for inventory management, sales forecasting, or strategic planning.

Optimized Resource Allocation: With accurate predictions, the business can optimize resource allocation more effectively. For instance, accurate demand forecasting can lead to better inventory management, reducing both overstock and stockouts.

Enhanced Customer Experience: The model's ability to predict customer behavior or demand accurately can improve customer satisfaction. By anticipating customer needs and preferences, the business can tailor its offerings, marketing strategies, and promotions more effectively.

Competitive Advantage: Utilizing a high-performing model like Gradient Boosting can give the business a competitive edge. It enables more precise and informed strategies compared to competitors who may rely on less accurate models.

Cost Efficiency: Accurate predictions reduce the risk of costly mistakes such as excess inventory or missed sales opportunities. This can lead to significant cost savings and improved profitability.

Overall, the Gradient Boosting model's superior performance makes it a valuable asset for the business, supporting more informed, data-driven decisions and contributing to operational efficiency and strategic success.

Additional Business Implications of the Optimum Model

Strategic Planning Enhancement:

Refines long-term strategies with precise forecasts.
Supports effective market expansion and product development.

Targeted Marketing Campaigns:

Enables highly targeted marketing efforts.
Improves engagement and conversion rates.
Optimizes marketing budget use and return on investment.

Supply Chain Optimization:

Enhances demand forecasting accuracy.
Improves coordination between production, distribution, and inventory.
Minimizes delays and boosts efficiency.

Risk Management:

Provides insights for proactive risk mitigation.
Helps address demand fluctuations and supply chain disruptions.

Customer Segmentation:

Identifies distinct customer segments.
Allows for personalized offers and recommendations.
Enhances customer satisfaction and loyalty.

Pricing Strategy Improvement:

Informs dynamic pricing strategies based on demand forecasts.

Maximizes revenue and market competitiveness.

Product Development:

Guides product development by highlighting trends and preferences.
Drives innovation and aligns products with market needs.

Financial Forecasting:

Enhances accuracy in financial forecasting.
Aids in budget planning, cash flow management, and achieving financial targets.

Operational Efficiency:

Reduces uncertainty in forecasts for streamlined operations.
Optimizes resource utilization and reduces waste.

Competitive Positioning:

Positions the business as an industry leader with advanced analytics.
Provides a competitive edge through better predictions and insights.

Conclusion:

The Gradient Boosting model provides the most reliable and accurate predictions for the business problem at hand. By leveraging this model, the FMCG company can significantly enhance its operational efficiency, reduce costs, and improve profitability through better demand forecasting and inventory management. This model's implementation will support strategic decision-making and drive business growth.