4/14/2024

Supply Chain

CAPSTONE PROJECT

C.SHRIRAJBHARETH
BATCH:PGPDSBA.O.APR23.A

TABLE OF CONTENT

LIST OF TABLES:

LIST OF FIGURES:

1.1 **PROBLEM STATEMENT**

- **The problem statement involves optimizing the supply chain for a Fast-Moving Consumer Goods (FMCG) company's instant noodles business.**
- **We've identified a significant discrepancy in demand and supply across our nationwide warehouses, resulting in elevated inventory costs.**
- **The primary objective is to determine the ideal product weight for shipments to each warehouse, aiming for a harmonious balance that meets high-demand areas efficiently while curbing unnecessary surplus in low-demand regions.**
- **This initiative aligns with our commitment to enhancing operational efficiency, reducing costs, and creating a more responsive and adaptive supply chain for our instant noodles business**.

1.2 NEED OF THE PROJECT

- **Addressing supply chain inefficiencies to cut inventory costs and enhance operational efficiency is imperative.**
- **Streamlining supply chain processes presents a vital opportunity for sustainability and profitability in the FMCG market.**
- **Optimizing supply quantities based on historical data enables efficient customer demand fulfillment, driving satisfaction and market share.**
- **Beyond business gains, such practices contribute to resource efficiency and potentially positive community impacts.**

- **Essential steps toward reducing costs, improving performance, and fostering sustainability in the competitive FMCG landscape.**

2) EXPLORATORY DATA ANALYSIS

Data Overview:

The imported data has 25000 rows and 24 columns

```
(25000, 24)
```

Duplicate Values :

There are no duplicate values present in the dataset.

Descriptive Details

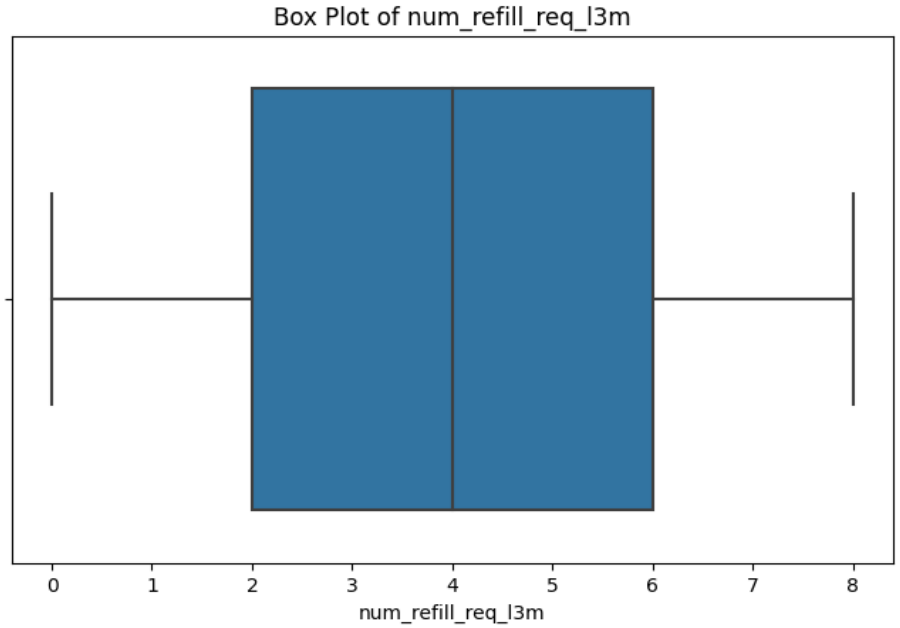
	count	mean	std	min	25%	50%	75%	max
num_refill_req_l3m	25000.0	4.089040	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	0.773680	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	3.104200	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	4985.711560	1052.825252	1821.0	4313.0	4859.0	5500.0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

The dataset contains 2 float, 14 integers, 8 objects

```
<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
```

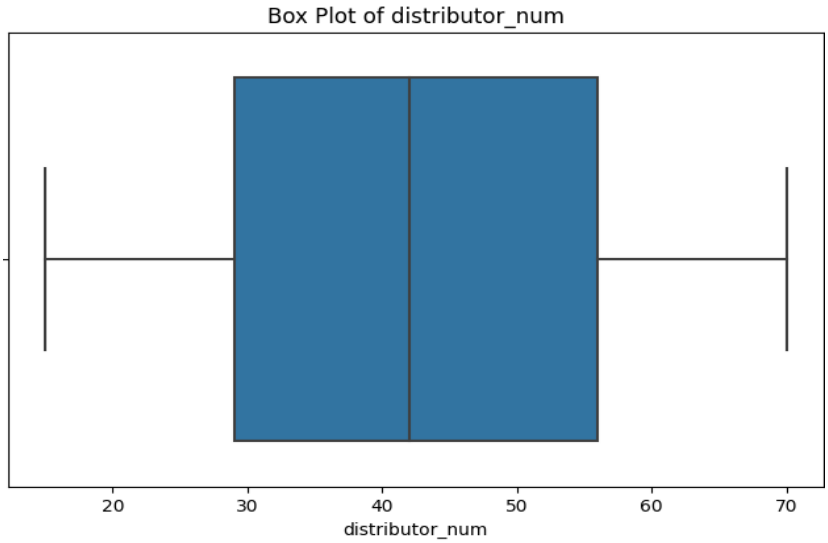
UNIVARIATE ANALYSIS:

Box Plot of num_refill_req_l3m



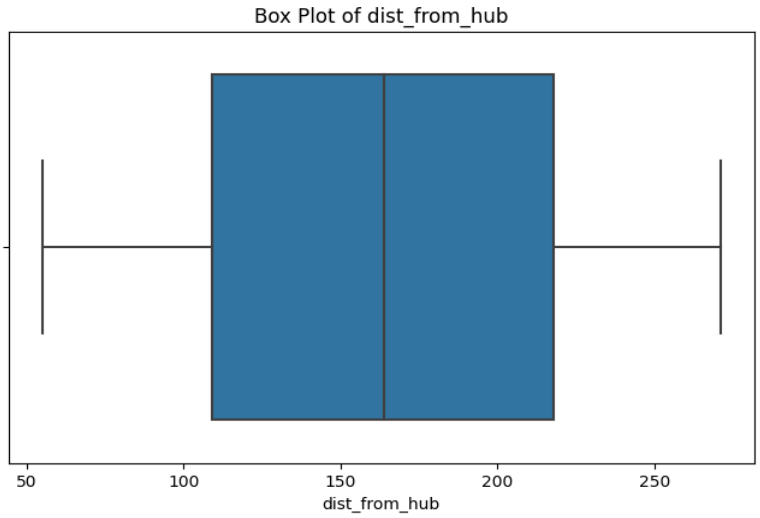
The box plot displays the median refill count, with the upper whisker indicating the maximum observed refill frequency of 8 times in the last 3 months.

Box Plot of distributor_num



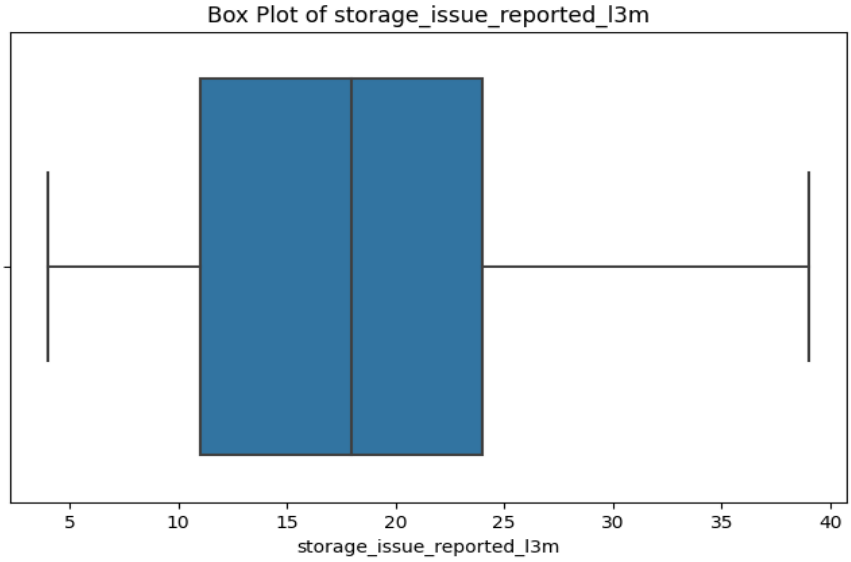
Median distributor works between warehouse and retail shops: 42; showcasing central tendency, with half of observed data points falling above and below this value.

Box Plot of dist_from_hub



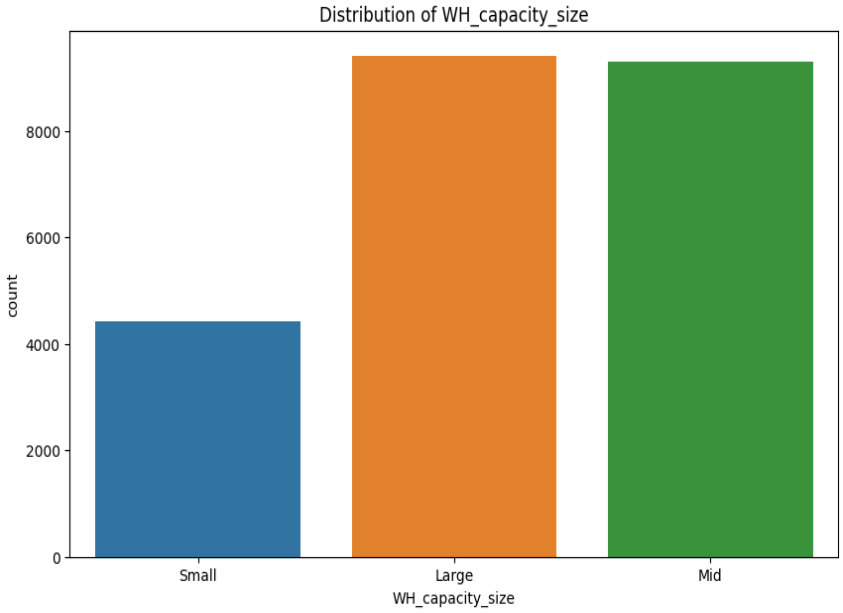
The central line in the box plot signifies that the median distance from the warehouse to the production hub is approximately 170 kilometers.

Box Plot of storage_issue_reported_l3m



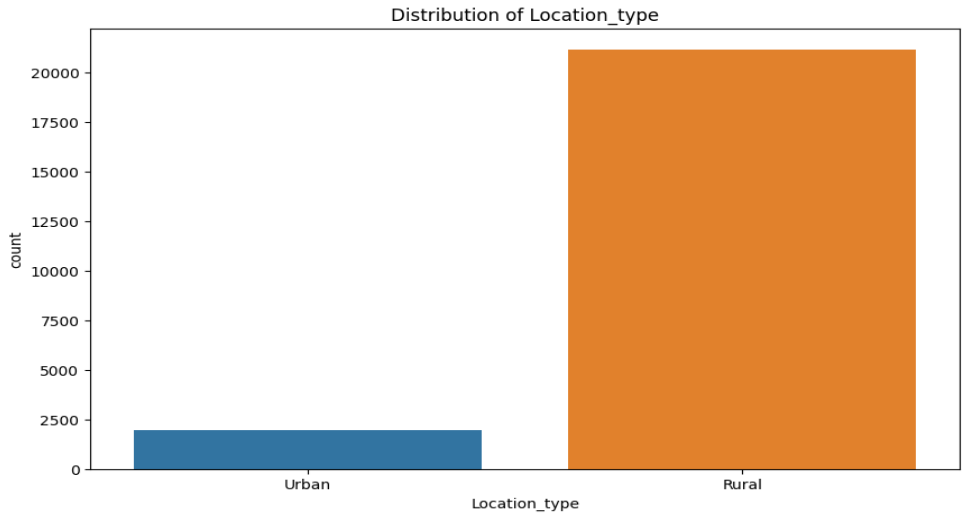
The central line in the box plot, positioned at 18, signifies the median frequency of storage issues reported to the corporate office.

Countplot of WH_capacity_size



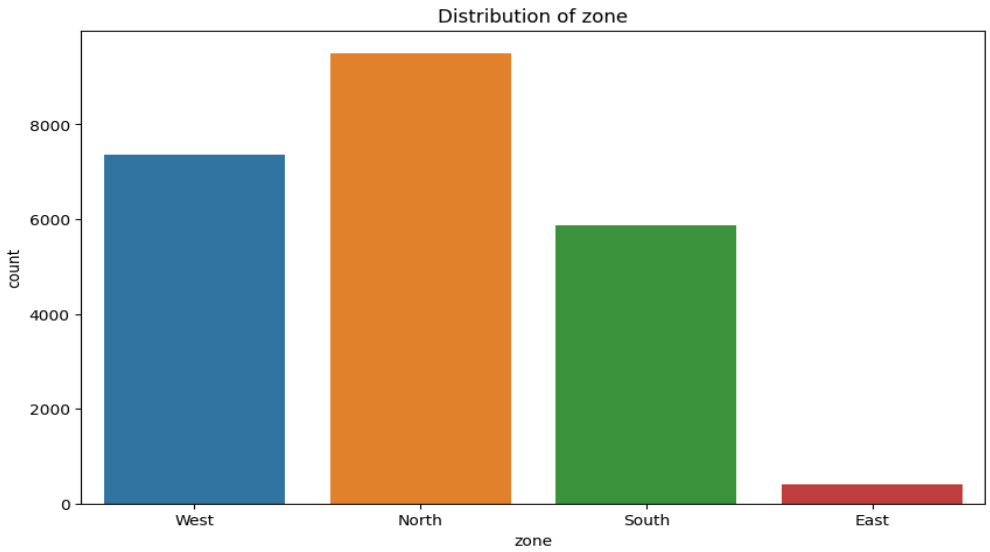
Warehouse capacity distribution: Majority have large capacity, followed by mid-sized and small; 10,169 large, 10,120 mid-sized, and 4,811 small capacity warehouses noted.

Countplot of Location_type



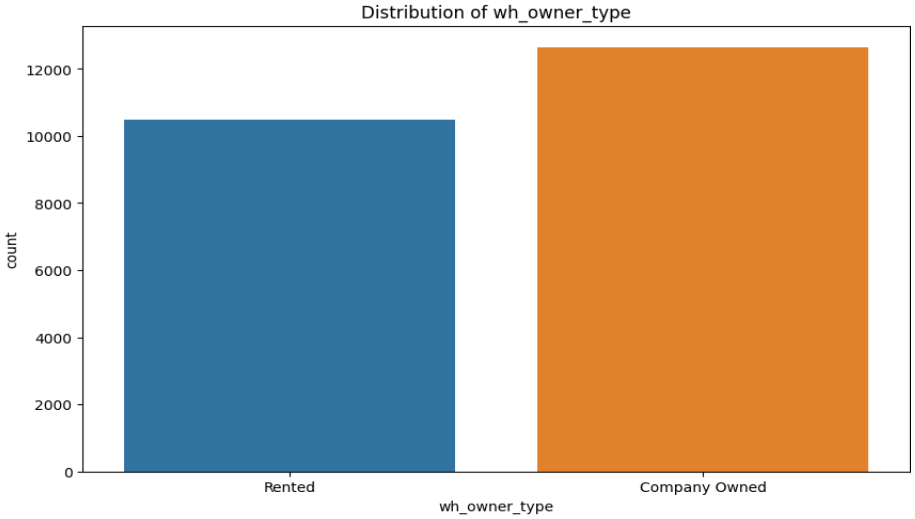
Warehouse distribution: Predominantly rural, with 22,957 urban and 2,043 rural warehouses identified.

Countplot of zone



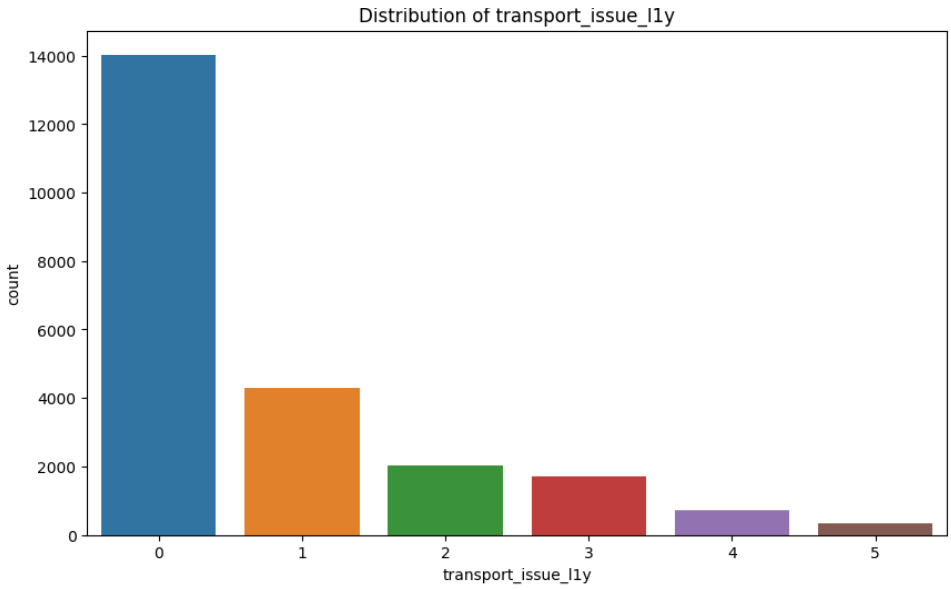
Warehouse distribution by zone: North zone leads with the highest count, followed by West and South; East zone notably lower in warehouse count.

Countplot of wh_owner_type

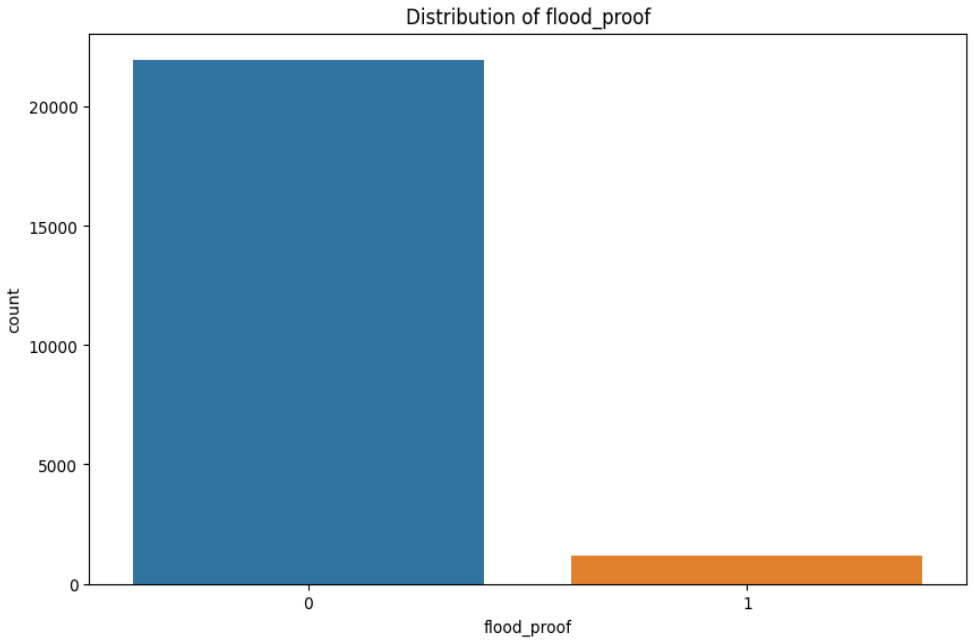


The information presented indicates that the company possesses a greater number of warehouses in comparison to the ones that are rented.

Countplot of transport_issue_l1y

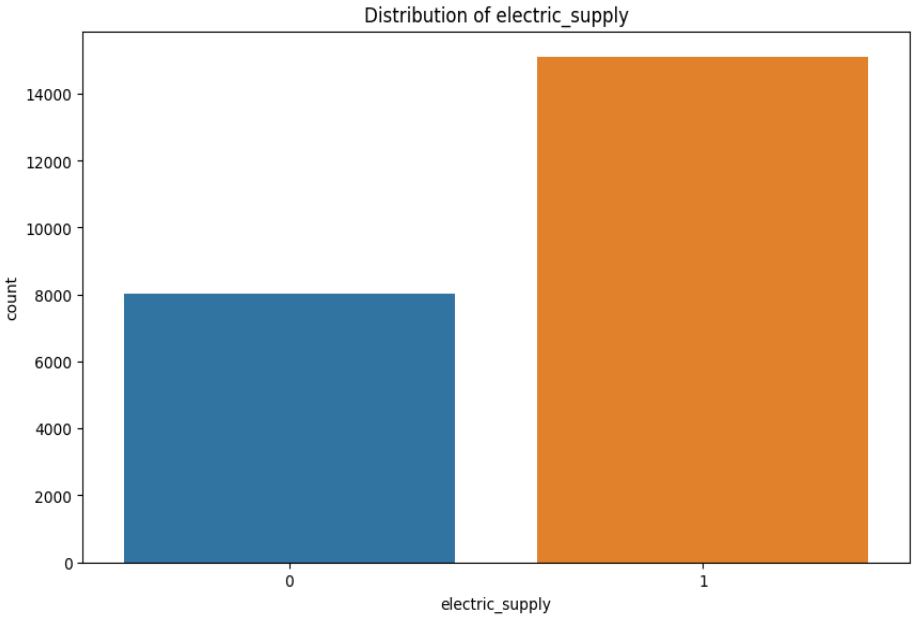


- **Reduced occurrence of transportation issues in the last twelve months observed.**
- **Highest count incident, occurring 5 times, observed 348 times in total.**



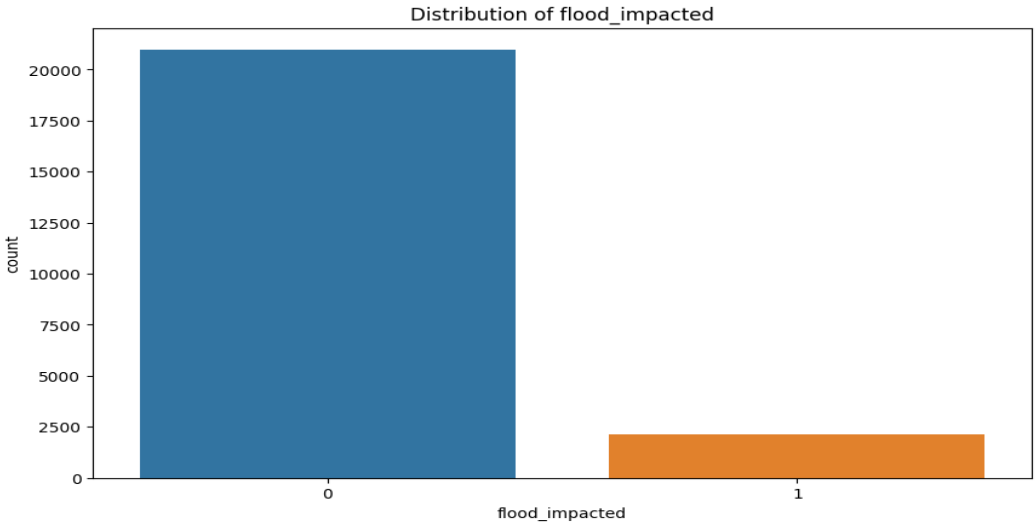
- **Warehouse elevation analysis: Binary variable indicates 0 for "True" (elevated) and 1 for "False" (ground level).**
- **Prevalence of 0 suggests most warehouses are elevated, aligning with observation of non-ground-level structures being common.**
- **Data visualization supports finding of elevated or non-ground-level warehouses as more prevalent in the dataset.**

Countplot of electric_supply



- **Visual representation indicates majority of warehouses equipped with electric backup systems, likely denoted by binary variable.**
- **Prevalence of backup systems ensures operational continuity during power disruptions, reflecting common and practical strategy among warehouses.**

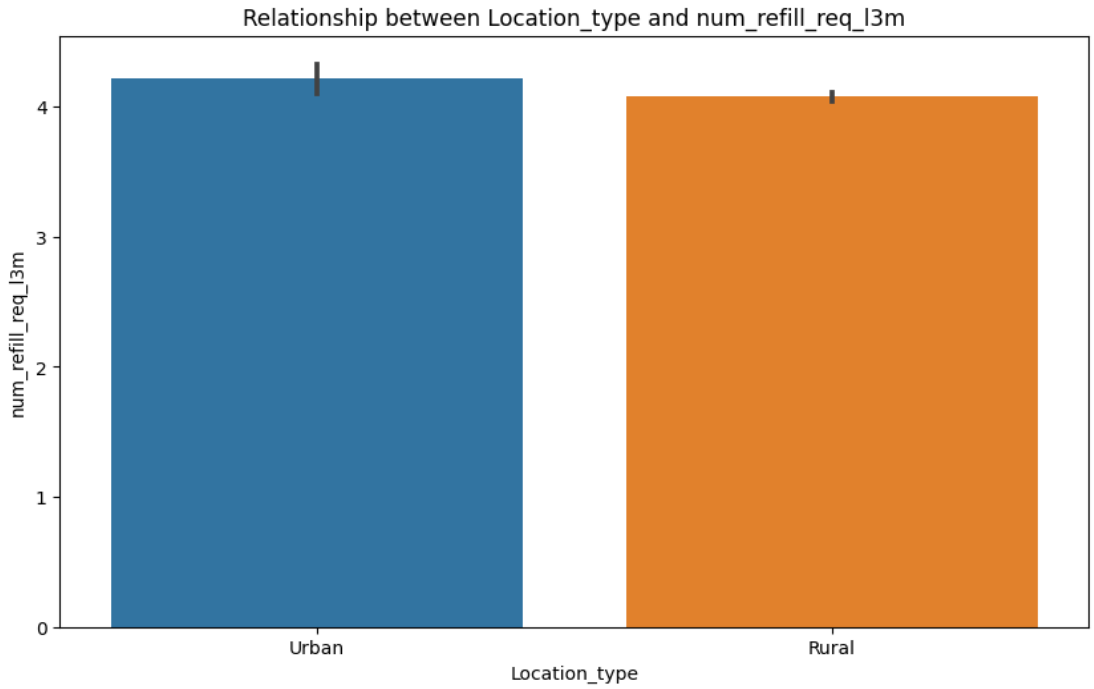
Countplot of flood_impacted



- **Countplot shows majority of warehouses not located in flood-impacted areas, likely indicated by binary variable.**
- **Prevalence of 0 suggests warehouses predominantly situated in regions unaffected by floods, indicating low susceptibility to flood impact.**

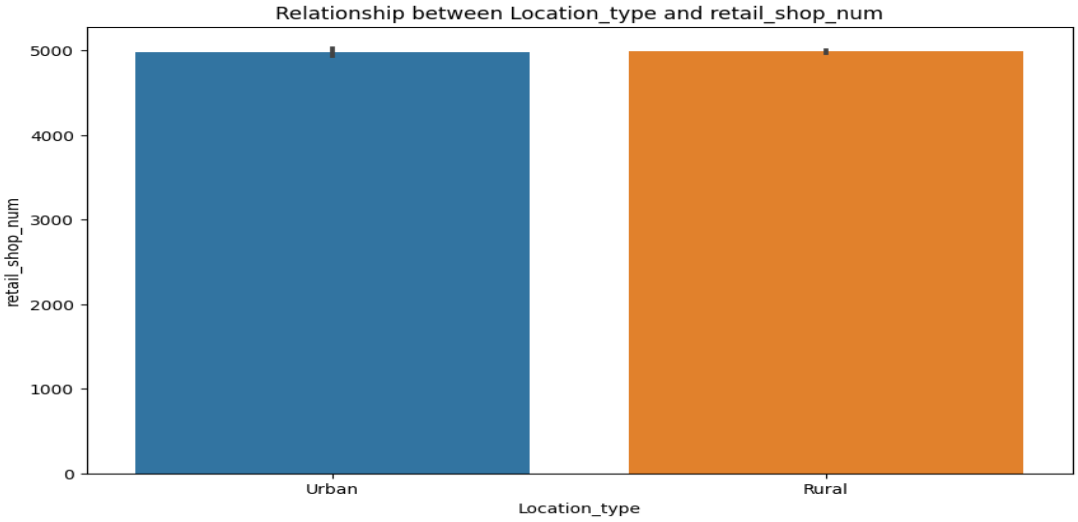
BIVARIATE ANALYSIS

Barplot showcasing relationship between Location_type and num_refill_req_l3m



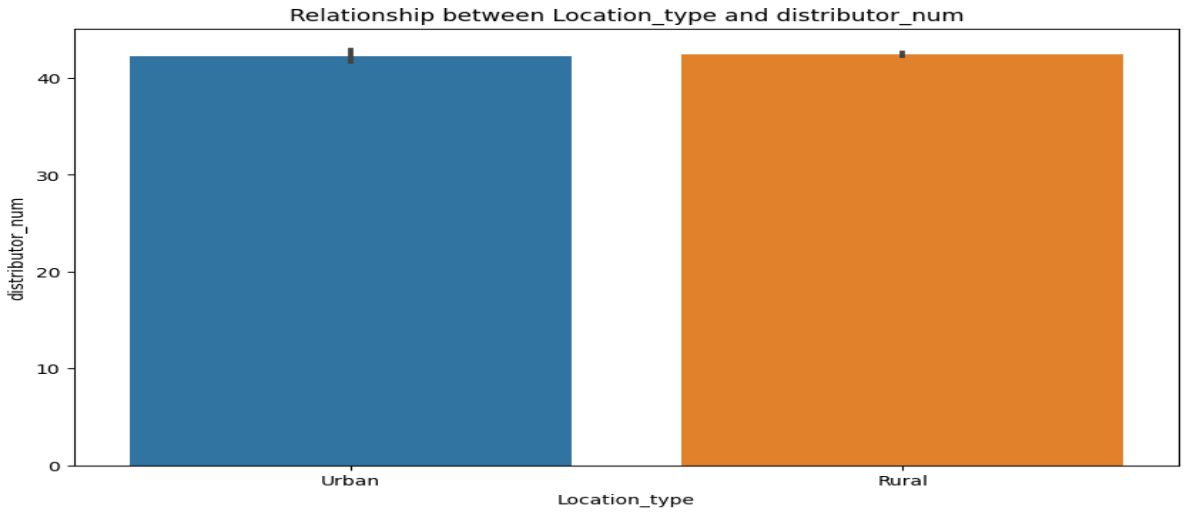
- **Location type strongly influences frequency of warehouse refilling activities in the last 3 months.**
- **Figure shows urban warehouses have notably higher refilling occurrence, indicating correlation between urban location and refilling frequency.**

Barplot showcasing relationship between Location_type and retail_shop_num



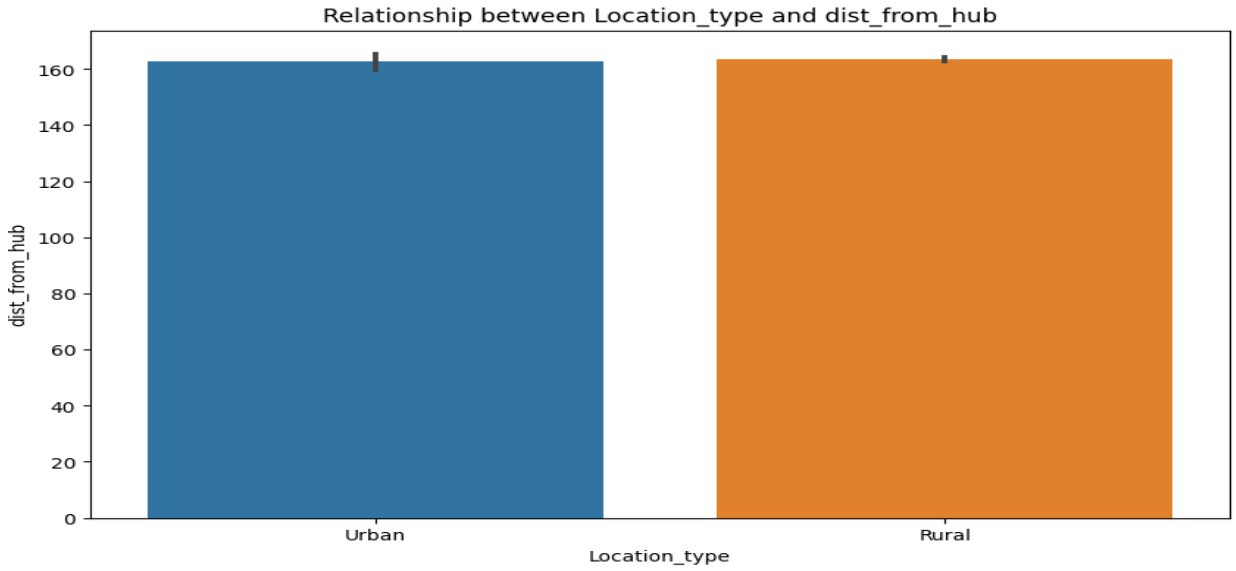
- **Figure shows urban and rural areas have similar counts of retail shops under warehouse area.**
- **Similarity suggests location type does not significantly influence presence of retail shops, implying other factors may be more influential.**

Barplot showcasing relationship between Location_type and distributor_num



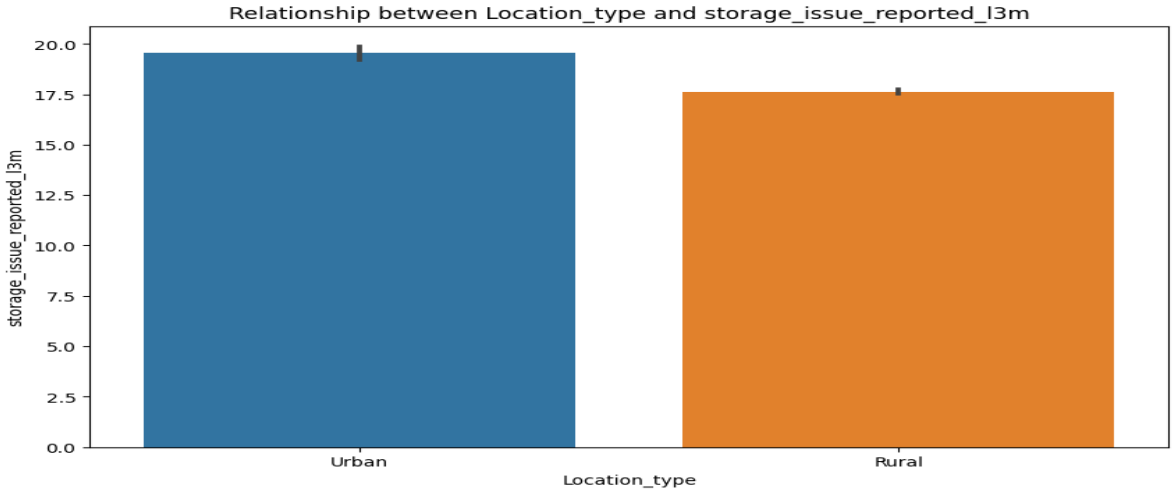
- **Plot shows no significant difference in distributor count between warehouse and retail shops in urban and rural areas.**
- **Similarity suggests location type does not substantially influence distributor engagement, indicating other factors may be more influential.**

Barplot showcasing relationship between Location_type and dist_from_hub



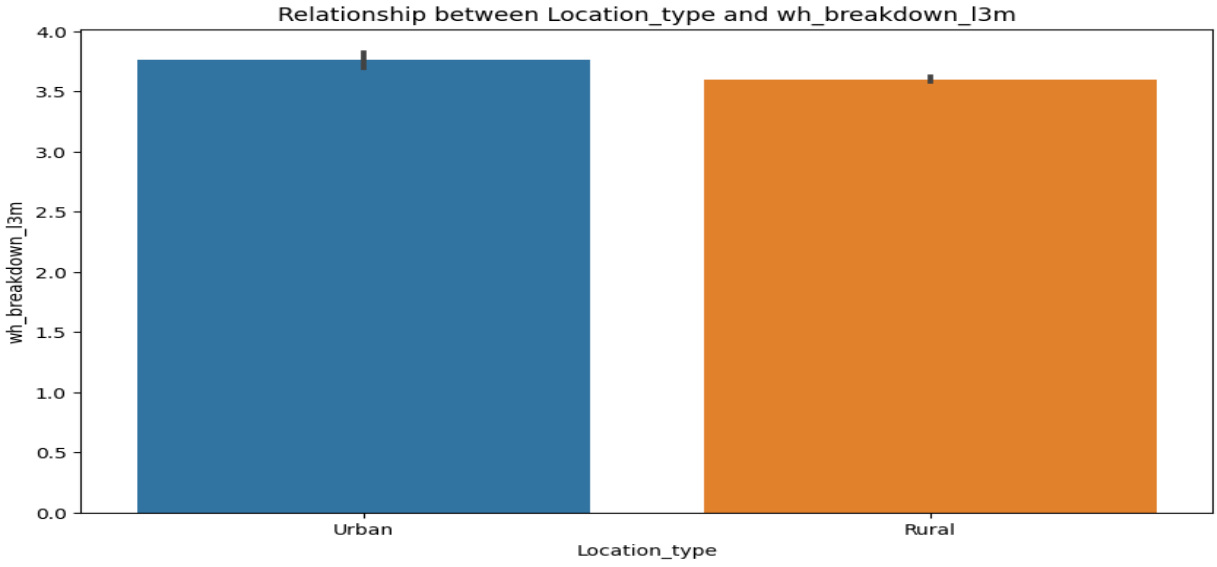
- **Plot shows average distance to production hub is similar for urban and rural locations.**
- **Similar distribution implies no significant difference in average distance, suggesting location type does not strongly influence proximity to production hub.**

Barplot showcasing relationship between Location_type and storage_issue_reported_l3m



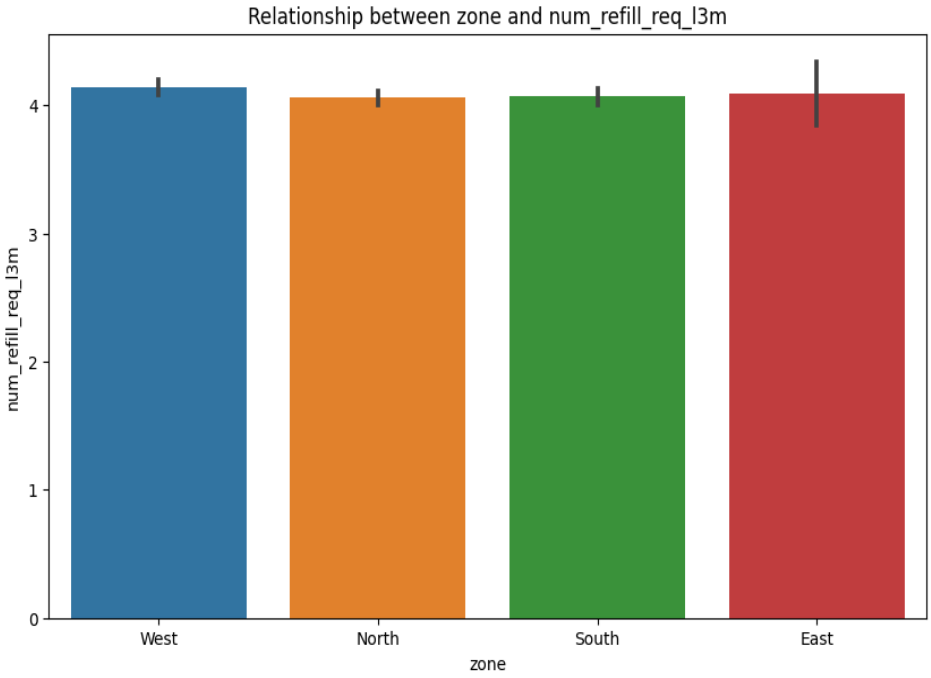
- **Barplot shows urban locations report higher number of warehouse storage issues in last 3 months.**
- **Contrastingly, rural areas report slightly lower count, suggesting minor role of location type in storage issues occurrence.**

Barplot showcasing relationship between Location_type and wh_breakdown_l3m



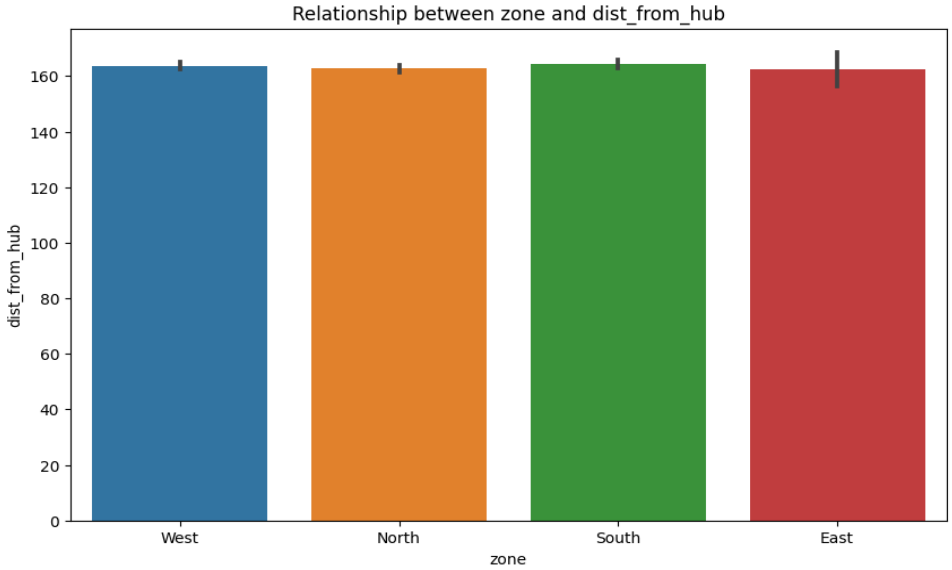
- **Plot illustrates more frequent warehouse breakdowns in urban areas, slightly lower count in rural areas.**
- **Lack of clear evidence suggests location type may not play major role in breakdown occurrences.**

Barplot showcasing relationship between zone and num_refill_req_l3m



- **Figure shows East zone has highest refilling count in last 3 months, followed by West and South.**
- **Lack of significant differences across zones suggests factors beyond geography may influence refilling frequency more prominently.**

Barplot showcasing relationship between zone and dist_from_hub



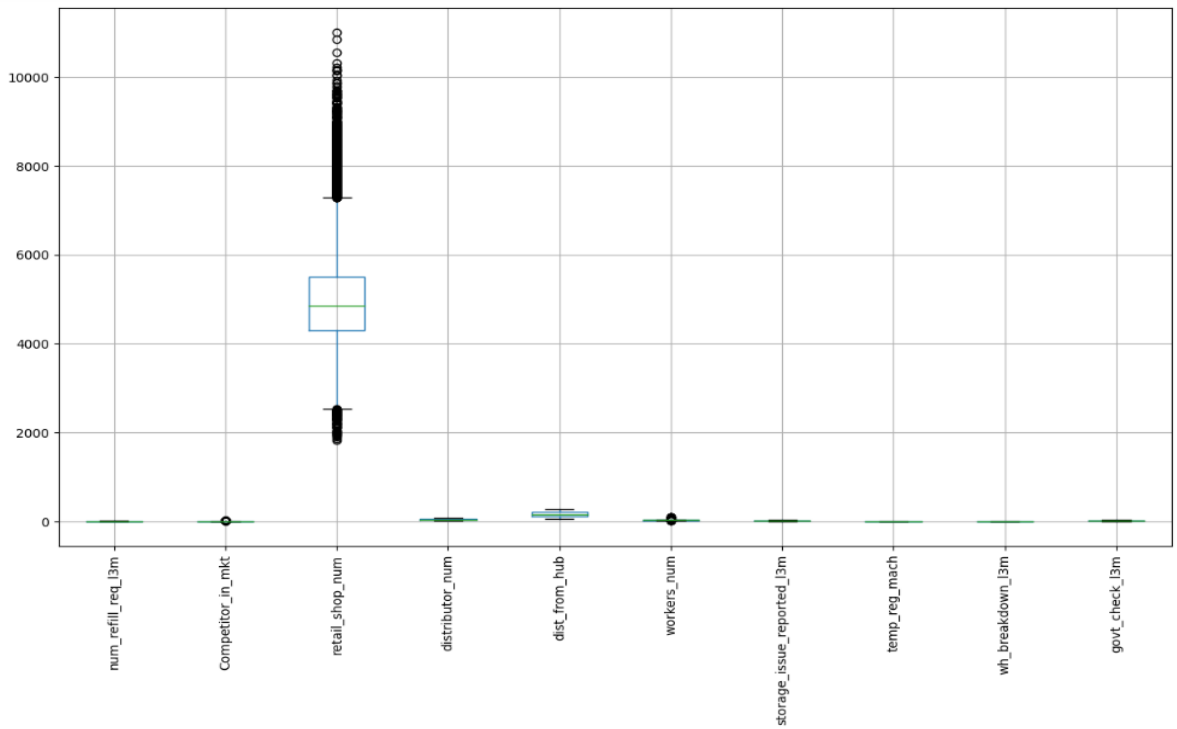
- **Plot shows East zone slightly farther from hub compared to other zones.**
- **West, North, and South zones have similar average distance of approximately 170 kilometers, indicating geographical distribution in relation to hub.**

3) DATA CLEANING AND PRE-PROCESSING

MISSING VALUE TREATMENT

- **"wh_est_year" column removed due to high count of null values, deemed impractical to impute or retain.**
- **For "workers_num" and "approved_wh_govt_certificate" columns with less than 4% null values, rows with missing values dropped using dropna() function to minimize impact on dataset integrity.**

OUTLIER TREATMENT



- **Outlier treatment not performed on dataset after careful consideration.**
- **Decision based on belief that outliers may hold valuable information or genuine variations in data, treating them could distort underlying patterns.**
- **By not conducting outlier treatment, aim is to maintain authenticity of data and allow for accurate representation of natural variability in features.**

VARIABLE TRANSFORMATION

- **No explicit variable transformation was performed**
- **The features used in the pipeline, including both numeric and categorical features, were processed using**

standardization and one-hot encoding techniques, respectively.

- **These preprocessing steps aimed to ensure compatibility of the features for subsequent feature selection and clustering algorithms.**

REMOVAL OF UNWANTED VARIABLES

```
Ware_house_ID                     0
WH_Manager_ID                     0
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
```

- **"Ware_house_ID" and "WH_Manager_ID" were excluded from the analysis as they were deemed unnecessary and unlikely to contribute significantly to the insights being sought.**
- **The variable "wh_est_year" was removed due to a high count of null values, which could compromise the reliability of insights derived from it.**

- **These exclusions aim to streamline the analysis by focusing on more relevant and complete data attributes, enhancing the reliability and relevance of the findings.**

4) MODEL BUILDING

- **The models used for building are Random Forest and Gradient Boosting.**
- **Random Forest can be employed descriptively to discern feature importance within a dataset.**
- **By assessing the significance of each feature, the model provides insights into the variables that most influence the target outcome.**
- **This aids in feature selection and understanding data intricacies.**
- **Similarly to Random Forest, Gradient Boosting can be utilized descriptively to ascertain feature importance.**
- **Through the identification of significant predictors and their impact on the target variable, Gradient Boosting indirectly offers prescriptive insights.**

Random Forest

```
Random Forest MSE: 762181.1766567754, R2: 0.9939768908633729
```

Gradient Boosting

```
Gradient Boosting MSE: 700215.6945461736, R2: 0.9944665708408987
```

- **The Gradient Boosting model outperforms the Random Forest model with a lower Mean Squared Error (MSE) of 700215.6945 compared to 762181.1767, indicating slightly better predictive accuracy.**
- **Both Random Forest and Gradient Boosting models demonstrate exceptional performance with high R-squared (R2) values, indicating their ability to explain a large proportion of the variance in the target variable.**
- **Gradient Boosting exhibits a slightly higher R2 value (0.9945) compared to Random Forest (0.9940), suggesting a marginally better fit to the data.**
- **Overall, both models exhibit strong performance on the test set, with minimal practical differences observed between them in terms of predictive accuracy and goodness of fit.**

Efforts to Improve Model Performance:

Feature Selection:

- **Analyze feature importance to identify influential variables and potentially refine the model by focusing on the most informative features.**
- **Competitor_in_mkt: This feature appears to have the highest importance (0.9833), indicating that it strongly influences the prediction of product weight. A higher number of competitors in the market may lead to adjustments in product weight to remain competitive.**
- **retail_shop_num: While having a lower importance compared to the competitor_in_mkt, the number of retail shops (0.0051) also plays a role in predicting product weight. More retail shops may**

indicate higher demand, impacting the quantity of product shipped.

Feature Importance

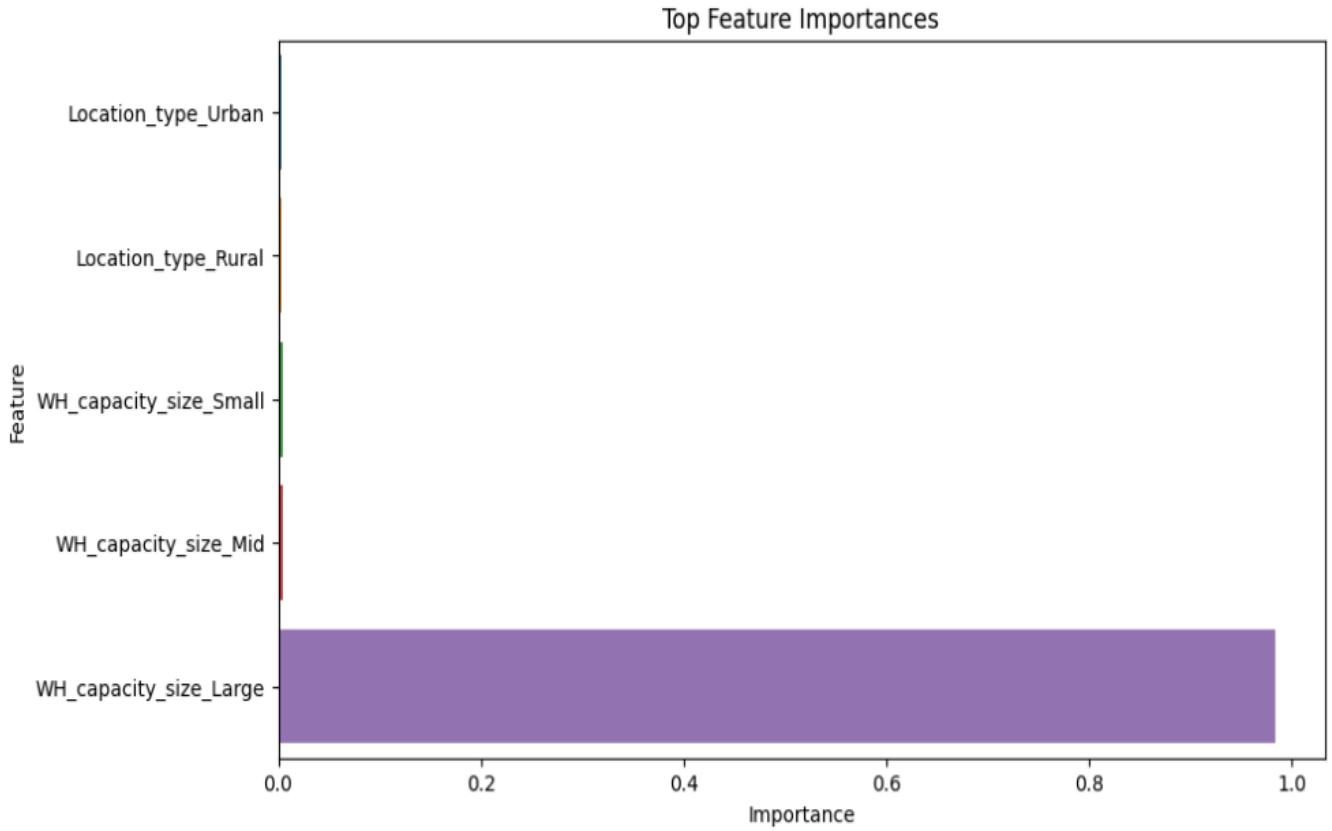
The top five most important features based on the feature ranking analysis are:

- **Number of competitors in the market (num__Competitor_in_mkt)**
- **Number of retail shops selling the product (num__retail_shop_num)**
- **Number of distributors working between the warehouse and retail shops (num__distributor_num)**
- **Number of times refilling has been requested in the last 3 months (num__num_refill_req_l3m)**
- **Presence of transport issues reported in the last year (num__transport_issue_l1y)**

These features provide valuable insights into the factors influencing product demand and supply within the market.

```
Feature ranking:
1. feature num__Competitor_in_mkt (0.9833284652536397)
2. feature num__retail_shop_num (0.005107629245105097)
3. feature num__distributor_num (0.004498365123033744 5)
4. feature num__num_refill_req_l3m (0.0038196645556274537)
5. feature num__transport_issue_l1y (0.0032458758225940372)
```

Feature Importance



5) MODEL VALIDATION

- **The model was validated by determining the most optimal choice through a comprehensive evaluation of various metrics, including Mean Squared Error (MSE), R-squared (R2), and business-specific considerations.**
- **Based on the provided MSE (Mean Squared Error) results, the Gradient Boosting model appears to be the most optimal choice.**
- **Lower MSE: Gradient Boosting has a lower MSE of 700215.69 compared to Random Forest's 762181.17. MSE measures the**

average squared difference between the predicted and actual values.

- **A lower MSE indicates a better fit for the data, signifying the model's predictions are on average closer to the actual values.**
- **Higher R-squared: While both models have very high R-squared values (close to 1), Gradient Boosting has a slightly higher R-squared (0.994467) compared to Random Forest (0.993977).**
- **R-squared represents the proportion of variance in the target variable that's explained by the model. A higher value suggests the model explains more of the variations in the data.**
- **Compared to Random Forest, Gradient Boosting offers better interpretability. We can extract feature importance from individual trees, providing insights into which features have the most significant impact on the model's predictions.**
- **As shown by the lower MSE in this case, Gradient Boosting can often achieve higher accuracy than Random Forest, especially for complex problems.**
- **In conclusion, based on the MSE and R-squared values, Gradient Boosting appears to be the most optimal model in this scenario. It offers a good balance between accuracy and interpretability.**

6) BUSINESS INTERPRETATION AND RECOMMENDATIONS

- **Explore establishing warehouses in diverse locations, including semi-urban areas, to tap into different market segments.**
- **Prioritize eco-friendly initiatives and renewable energy sources for warehouses to align with environmental responsibility and reduce operational costs.**
- **Implement advanced systems to minimize storage issues, anticipate demand, and streamline supply chain processes.**

- **Develop contingency plans and invest in supply chain resilience measures to mitigate risks associated with unforeseen events.**
- **Explore innovations such as customized packaging or loyalty programs to increase customer satisfaction and brand loyalty.**
- **Embrace IoT devices, automation, and data analytics to optimize warehouse operations, reduce costs, and improve productivity.**
- **Forge partnerships with local authorities to ensure compliance with regulations, leverage local resources, and enhance community engagement.**
- **Prioritize ongoing training programs for warehouse staff to stay updated on technological advancements and efficient workflow practices.**
- **Adopt lean principles to identify and eliminate inefficiencies, reduce waste, and optimize resource utilization in the production process.**
- **Explore dynamic pricing based on demand fluctuations, seasonal trends, and regional preferences to maximize revenue and competitiveness.**
- **Utilize Gradient Boosting for its lower MSE, leading to more accurate predictions compared to Random Forest.**
- **Leverage Gradient Boosting's feature importance capability to highlight factors with the greatest influence on the target variable.**
- **Acknowledge that Gradient Boosting models may be more complex but offer a compelling combination of accuracy, interpretability, and identification of key drivers, leading to better decision-making and business performance.**