Capstone Project

Supply Chain Project

By:

Name: Aishwariya Hariharan

PGP-DSBA Online September' 23

Date: 13 September 2024

Contents

Introduction	3
Exploratory Data Analysis	4
Data Cleaning and Preprocessing	35
Business Insights from Exploratory Data Analysis	39
Model Building and Model Validation	44
Interpretation of the Most Optimum Model	50
Business Implications	50
Recommendations	51

Introduction

Problem statement

An FMCG company entered into the instant noodles business two years back. Their higher management has noticed a mismatch in the demand and supply. Where the demand is high, supply is pretty low, and supply is quite high where the market is low. In both ways it is an inventory cost loss to the company; hence, the higher management wants to optimize the supply quantity in every warehouse in the entire country.

Need of the study/project

This exercise aims to build a model, using historical data that will determine the optimum weight of the product to be shipped each time to the warehouse.

Demand planning and supply chain management are crucial for optimising inventory levels, reducing costs, and improving customer satisfaction in the FMCG sector. By leveraging advanced forecasting and optimisation techniques, companies can effectively match supply with demand, enhance operational efficiency, and drive strategic decision-making. For the FMCG company in question, developing a robust model using historical data will provide a strong foundation for achieving these goals and demonstrating tangible impact to unlock further data resources and opportunities.

Understanding business/social opportunity

To analyse the demand pattern in different pockets of the country so management can drive the advertisement campaign, particularly in those pockets.

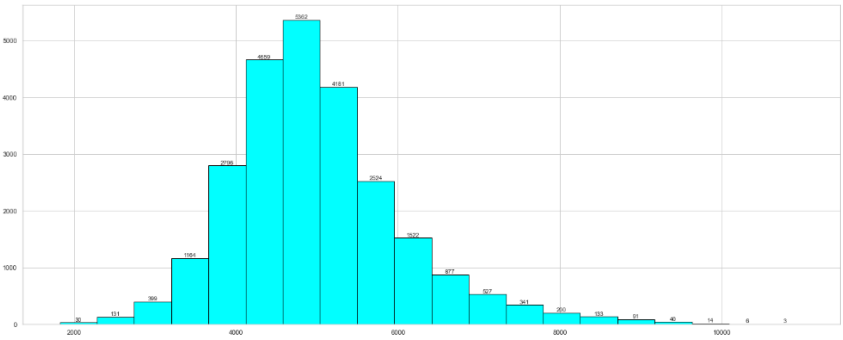
This is the first phase of the agreement; hence, the company has shared very limited information. Once you can showcase a tangible impact with this much information then the company will open the 360-degree data lake for your consulting company to build a more robust model.

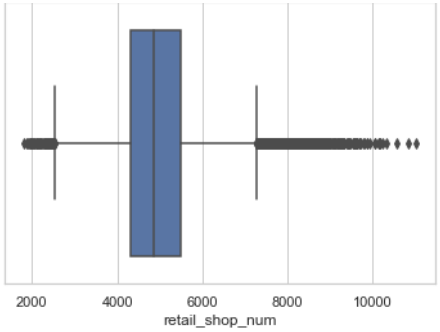
EDA and Business Implications

Univariate analysis

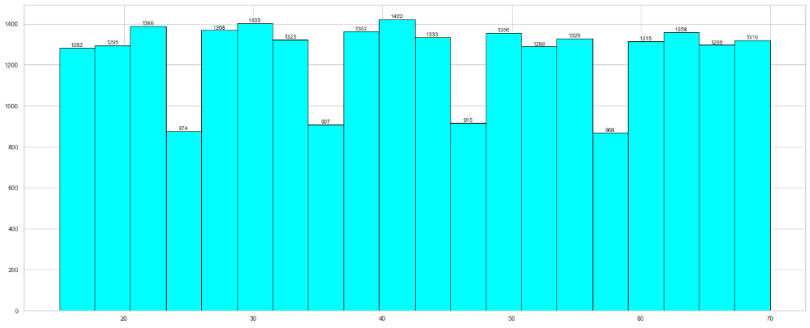
Distribution and spread of continuous attributes

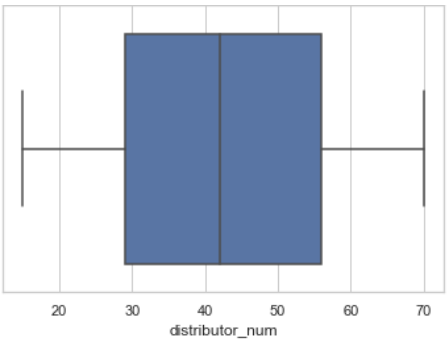
- **The number of retail shops that sell the product under the warehouse area:**



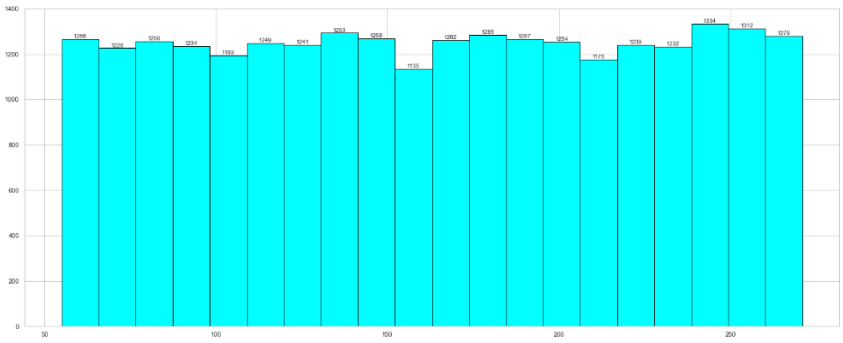


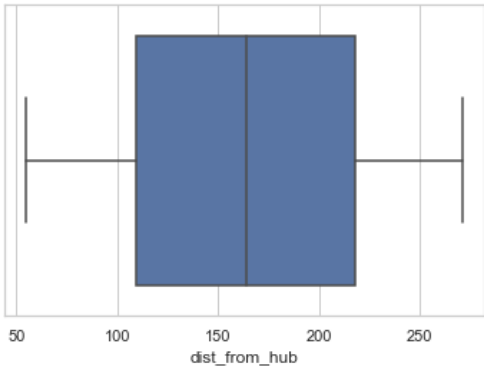
- **Number of distributer works in between warehouse and retail shops**



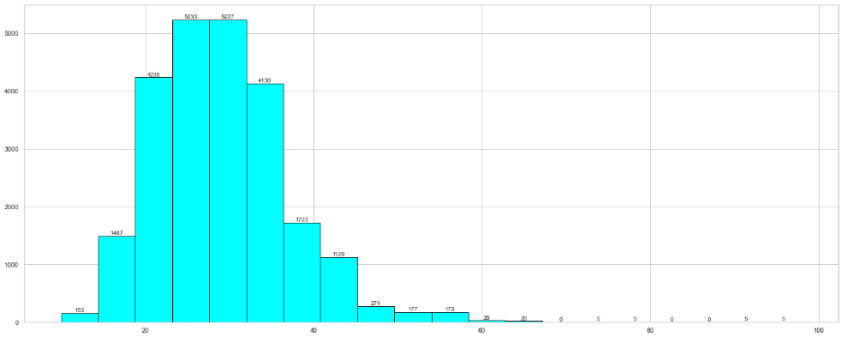


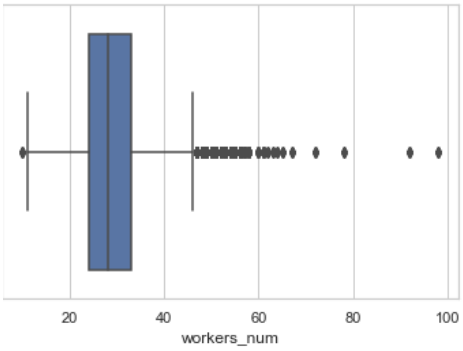
- **Distance between warehouse to the production hub in Kms**



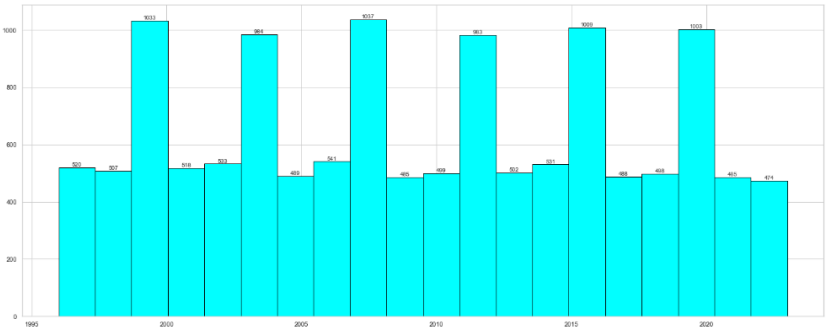


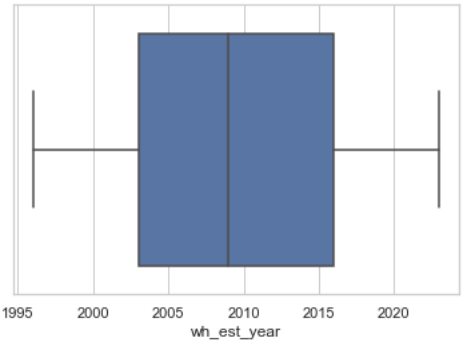
- **Number of workers working in the warehouse**



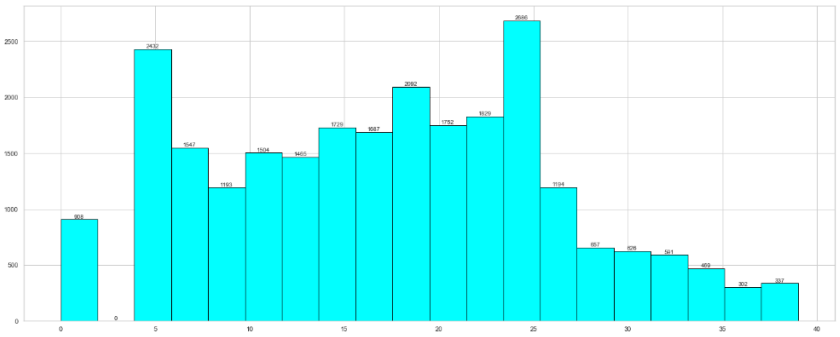


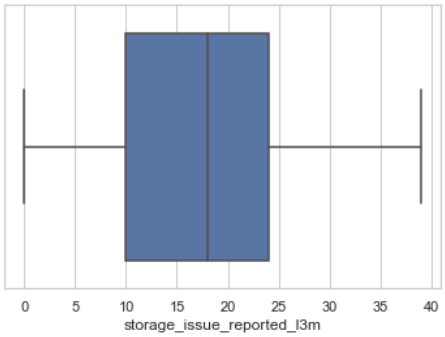
- **Warehouse established year**



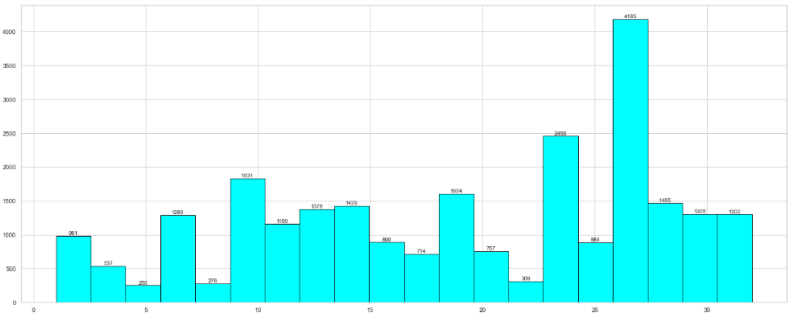


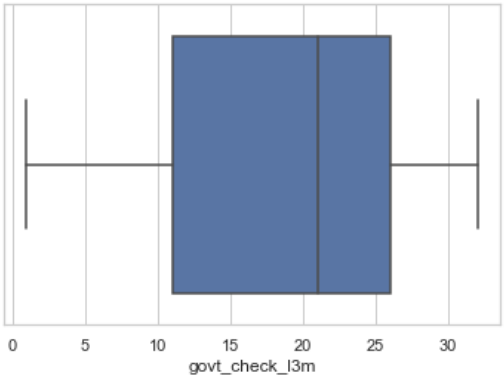
- **The warehouse reported storage issues to the corporate office in the last 3 months. Like rats, fungus because of moisture etc.**



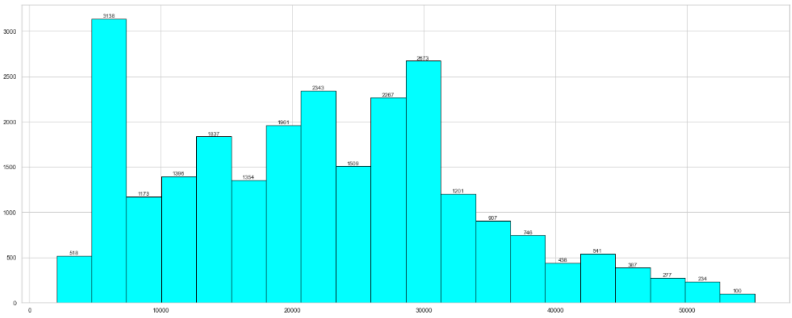


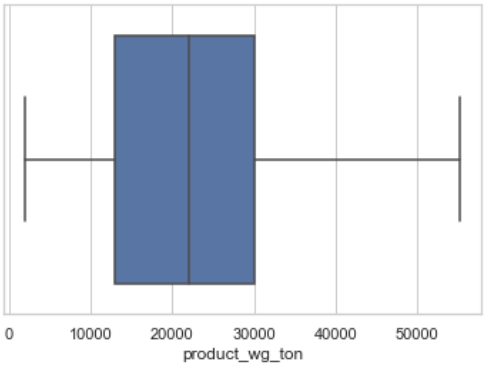
- **Number of times government officers have visited the warehouse to check the quality and expiry of stored food in the last 3 months**



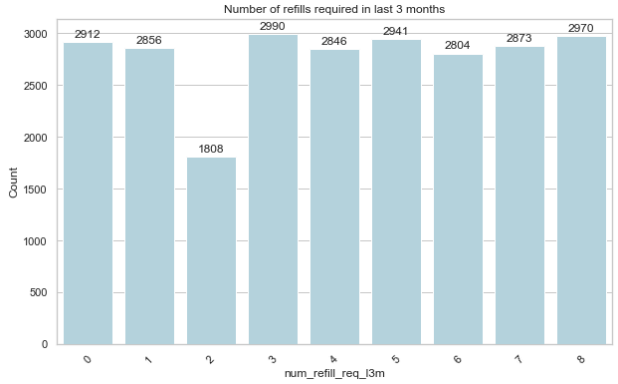


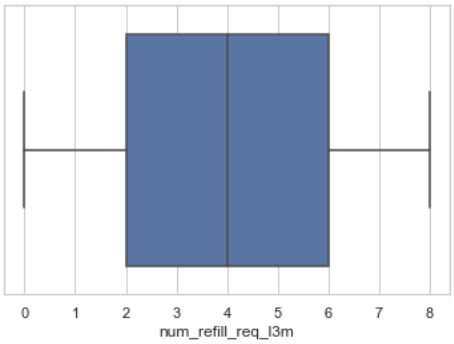
- **Weight (in tons) of the product that has been shipped in the last 3 months**



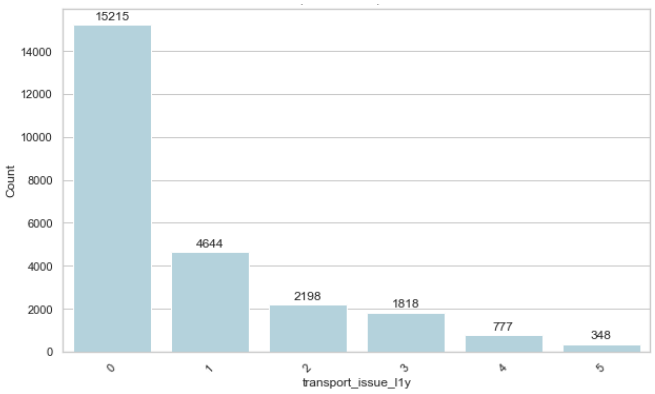


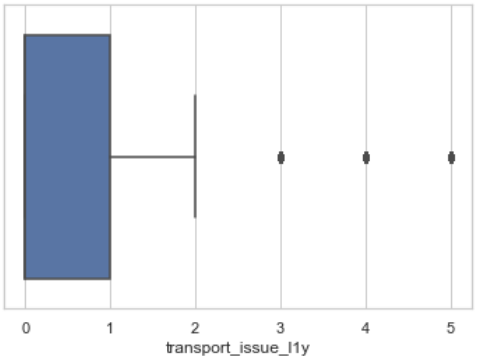
- **Number of times refilling has been done in last 3 months**



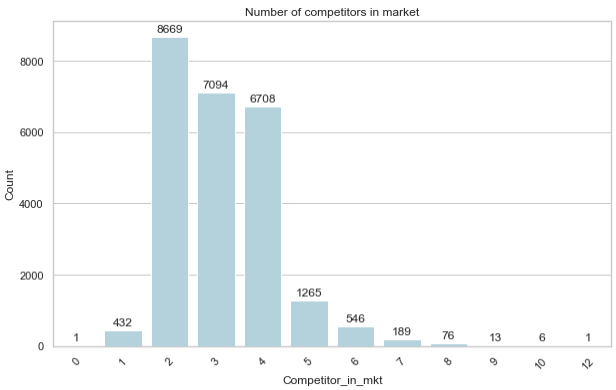


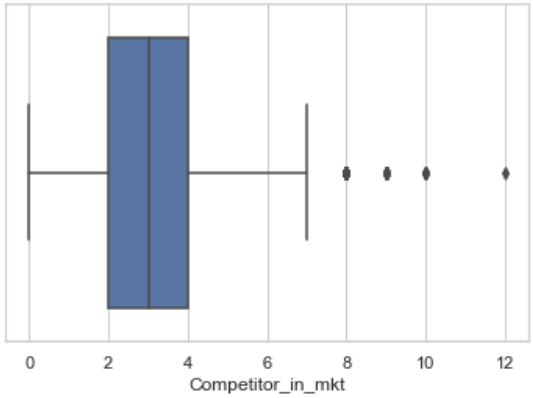
- **Number of transport issues like accidents or goods stolen reported in last year**



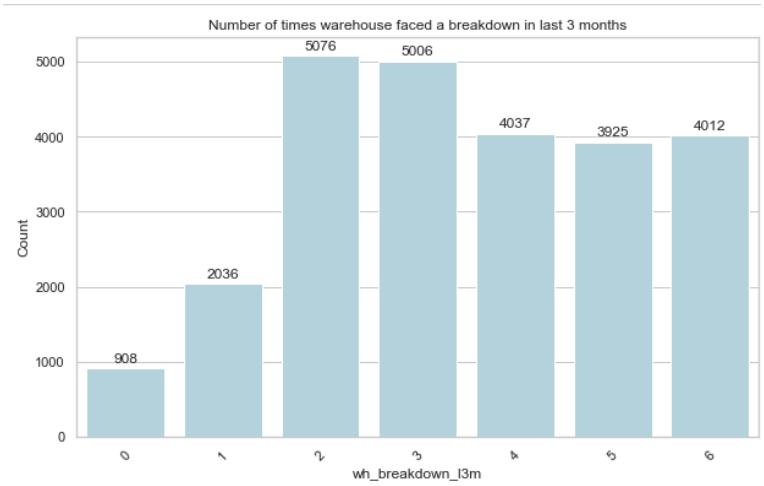


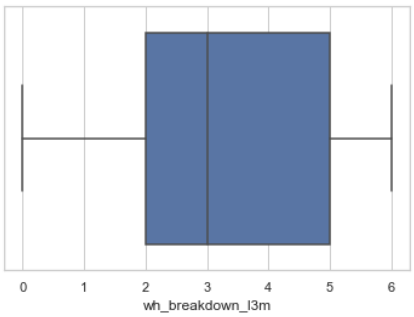
- **Number of instant noodles competitors in the market**





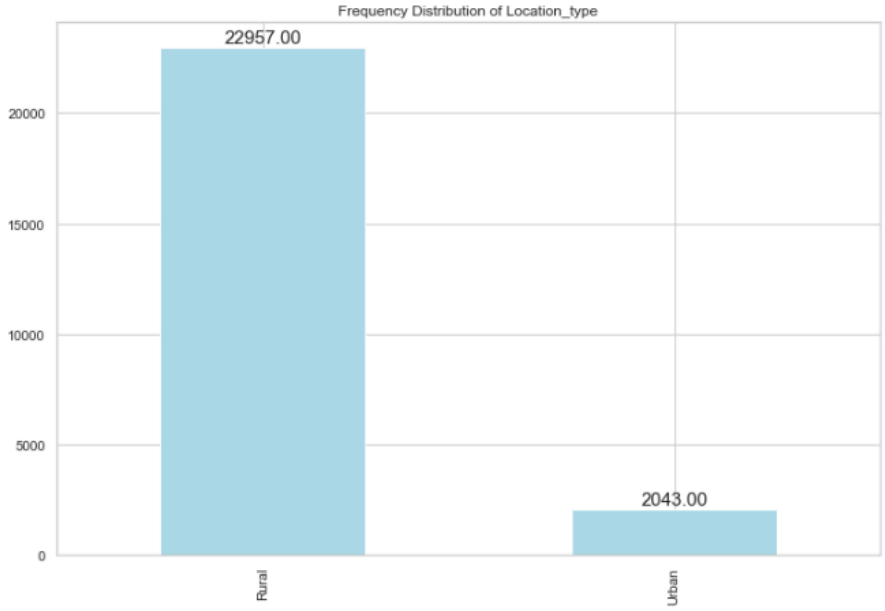
- **Number of times warehouse face a breakdown in last 3 months. Like strikes from workers, floods, or electrical failure**



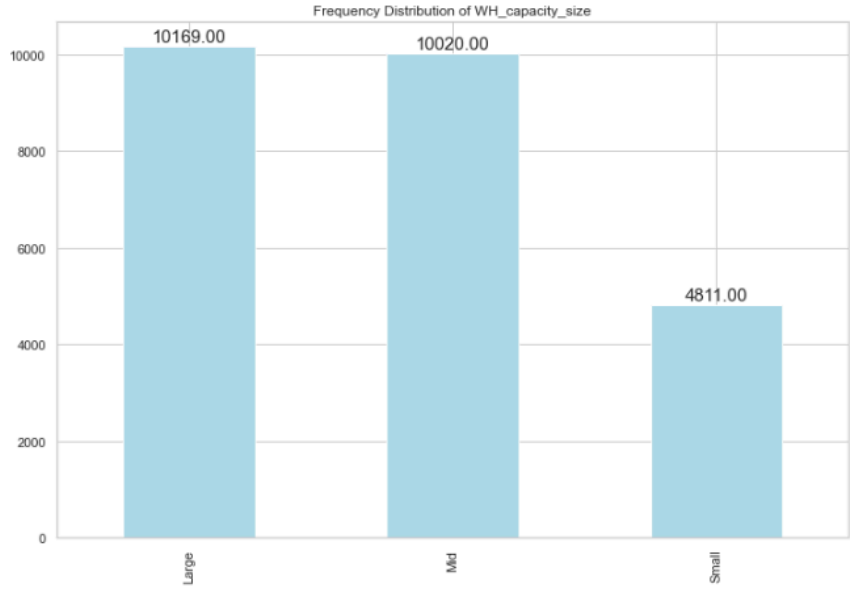


Distribution and spread of categorical attributes

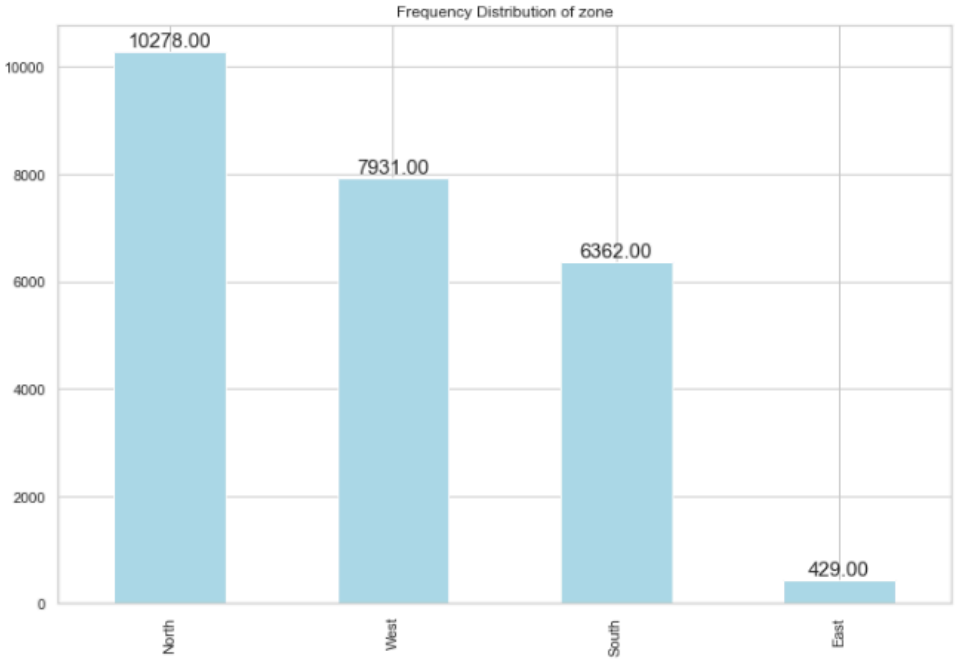
- **Frequency distribution of location types**



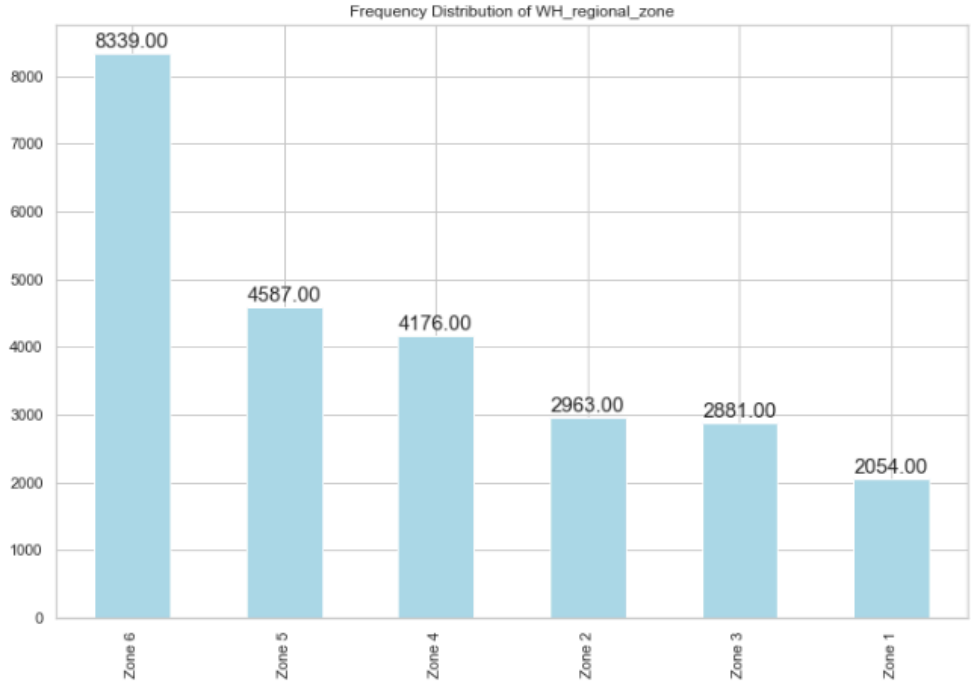
- **Storage capacity size of the warehouse**



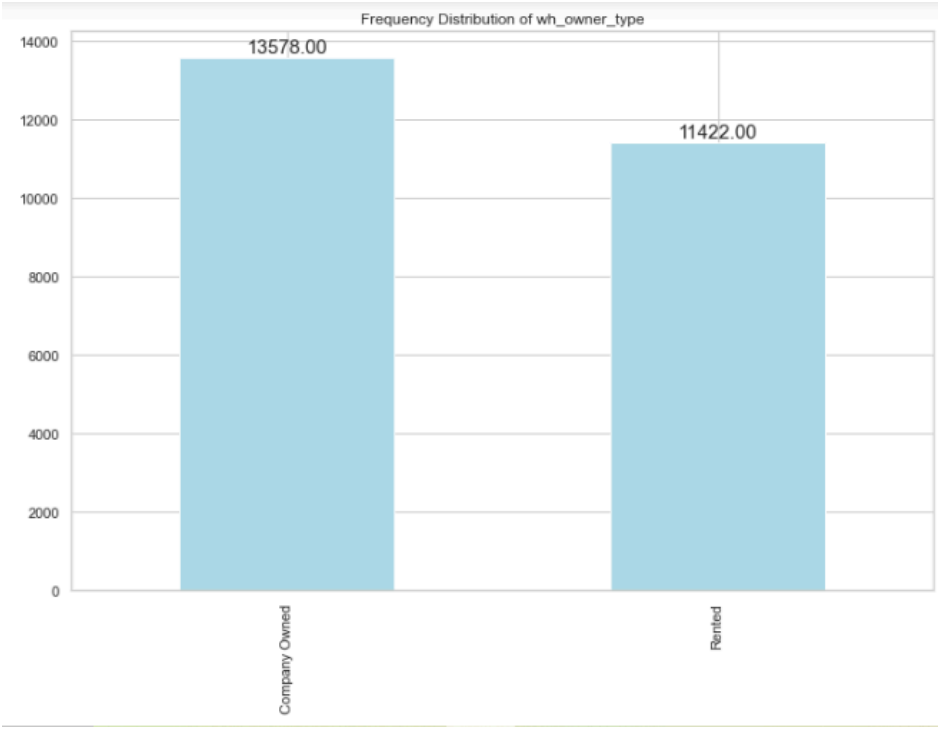
- **Frequency distribution of the zones of the warehouse**



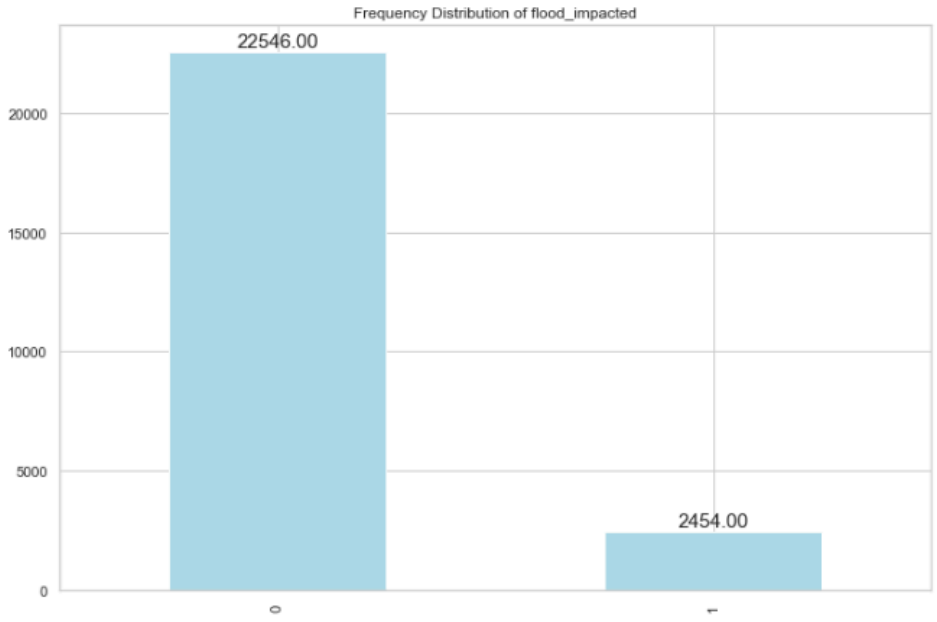
- **Regional zone of the warehouse under each zone**



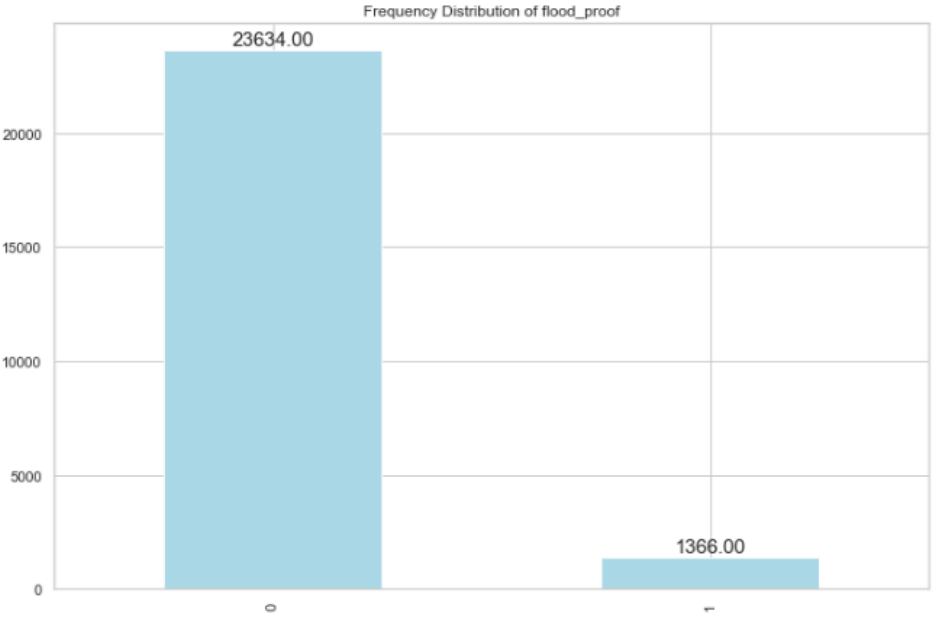
- **Frequency distribution of owner type**



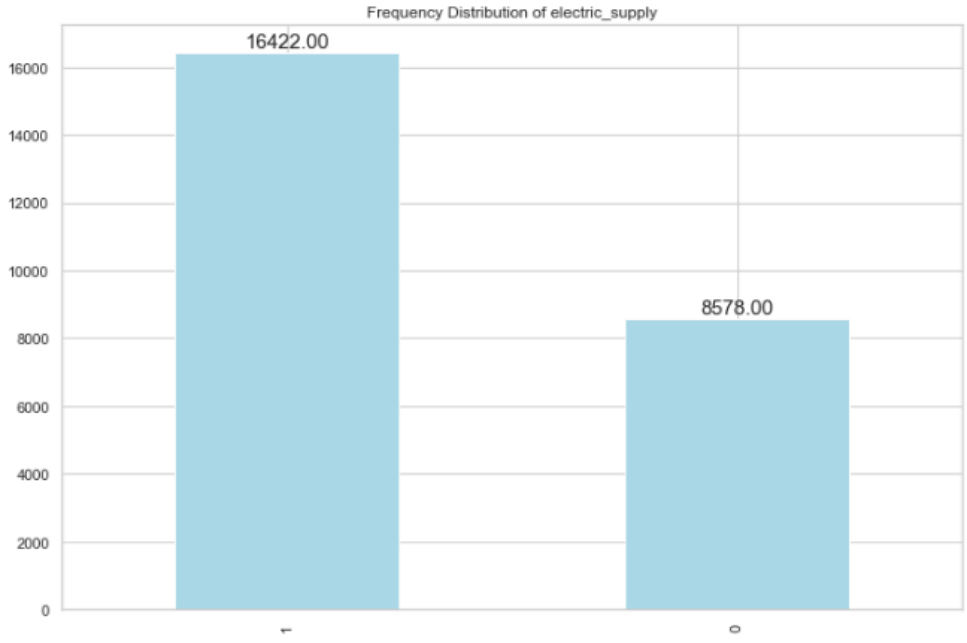
- **Indicator of whether the warehouse is in the flood-impacted area or not**



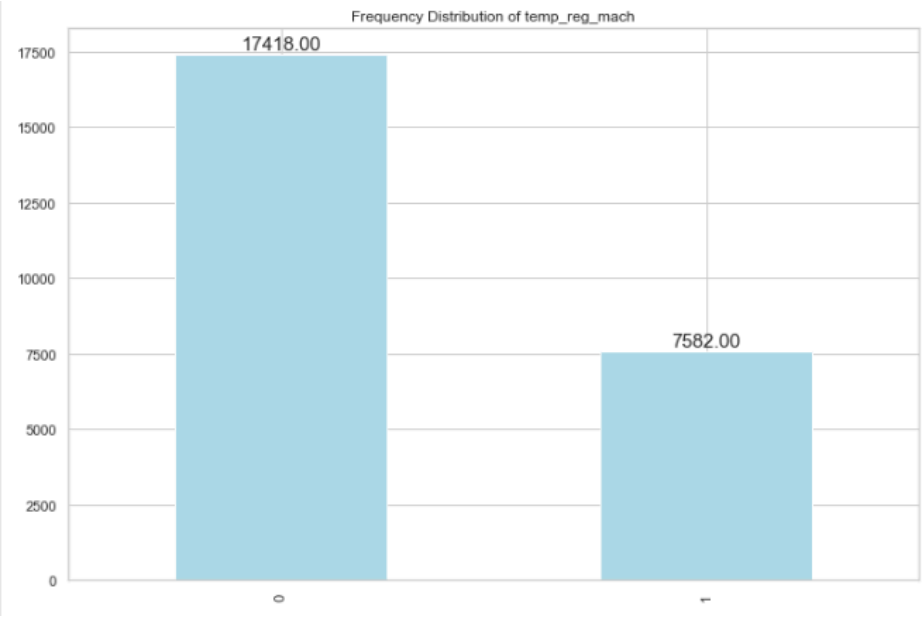
- **Indicator of whether the warehouse is in the flood-proofed or not**



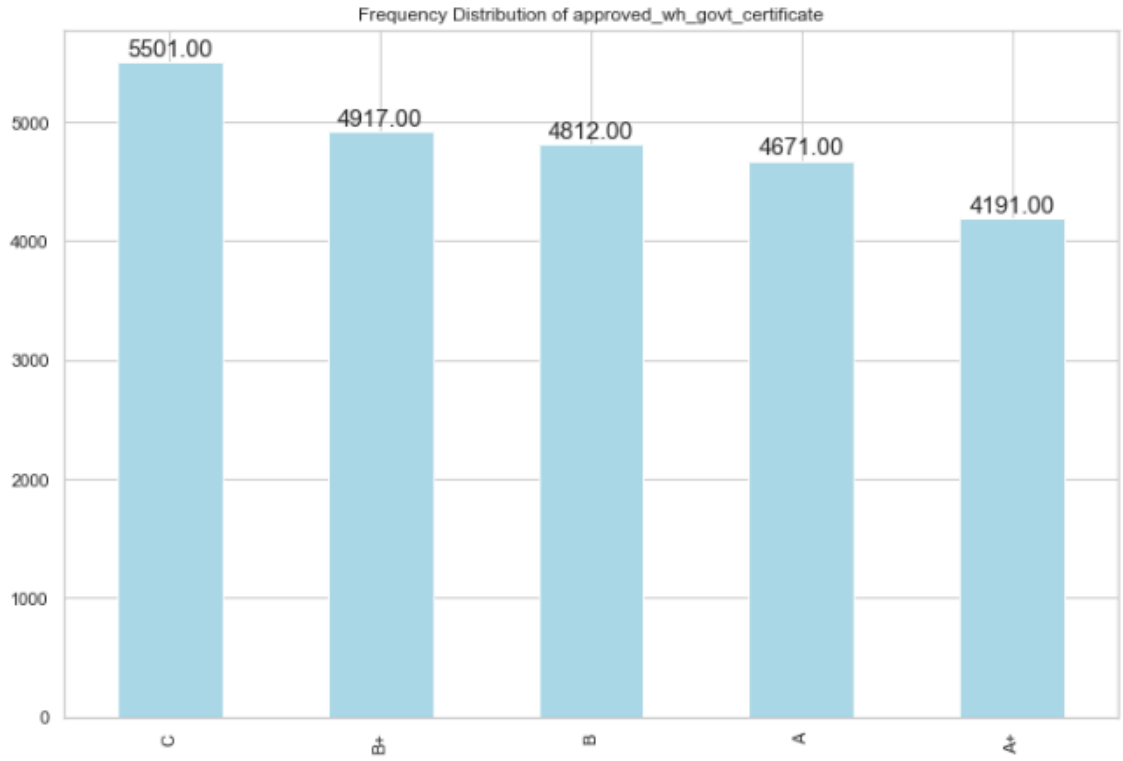
- **Indicator of whether the warehouse has electricity backup or not**



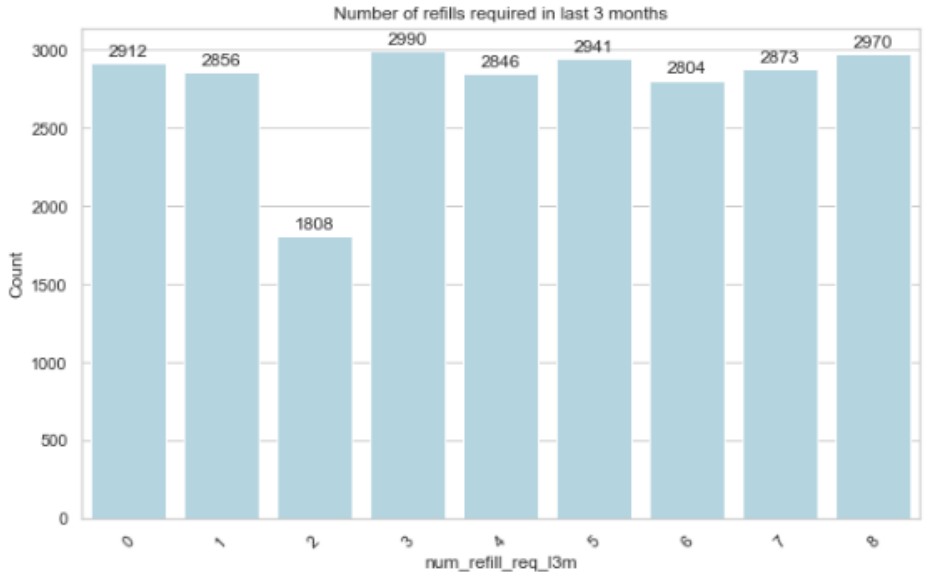
- **Indicator of whether the warehouse has a temperature regulating machine or not**



- **Type of standard certificate issued to the warehouse from a government regulatory body**

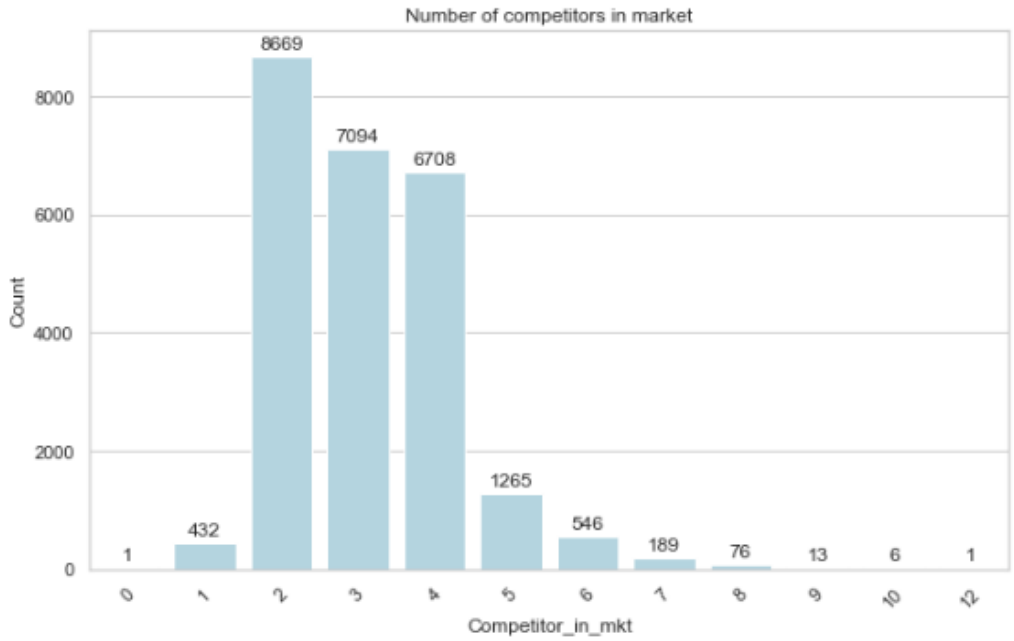


- **Number of refills required in last 3 months**

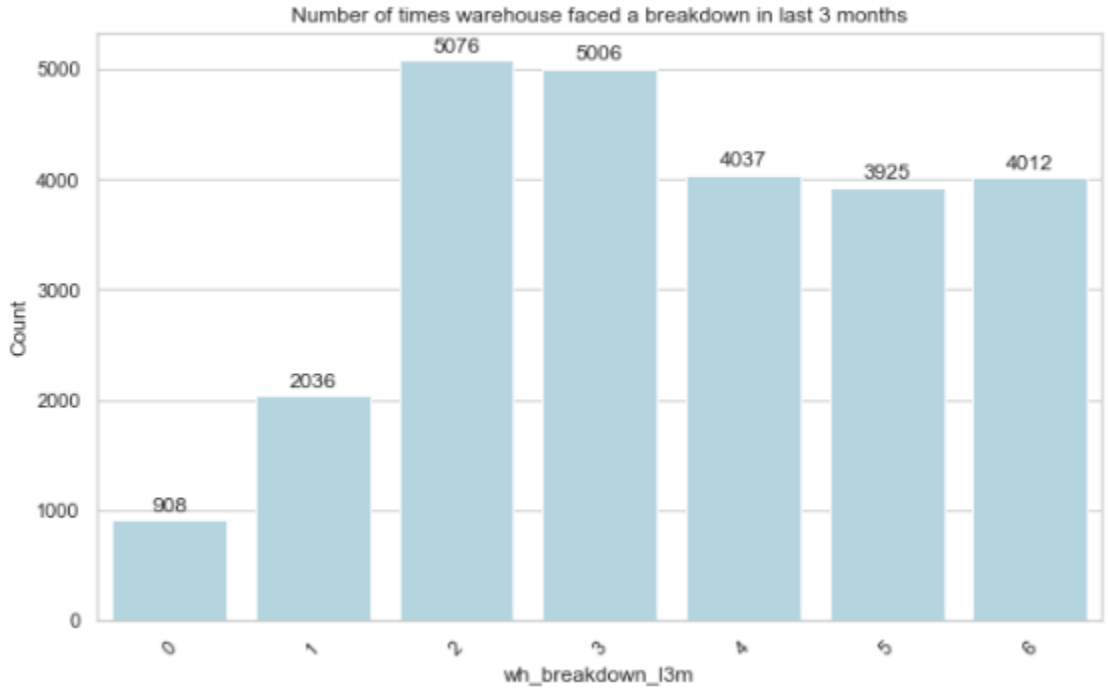


- **Number of transport issues reported in last 3 months**

- **Distribution of the number of competitors in the market**

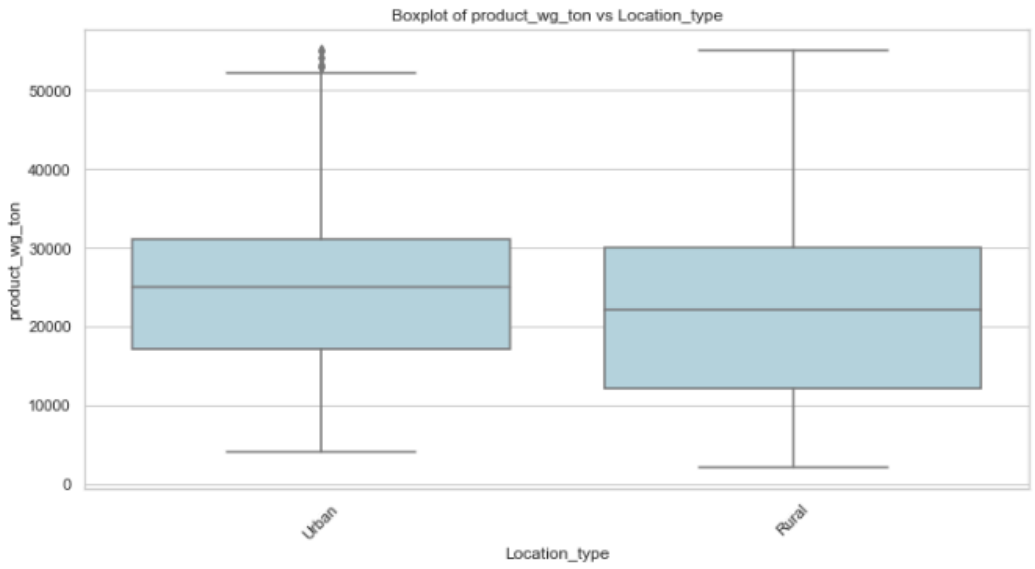


- **Distribution of the number of times warehouses faced a breakdown in last 3 months**

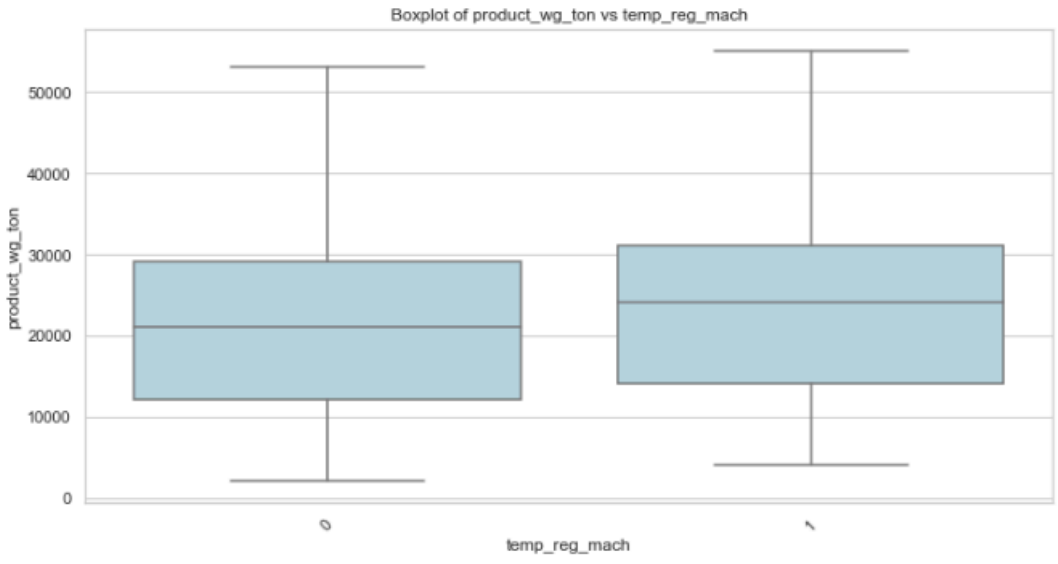


b) Bivariate analysis (relationship between different variables, correlations)

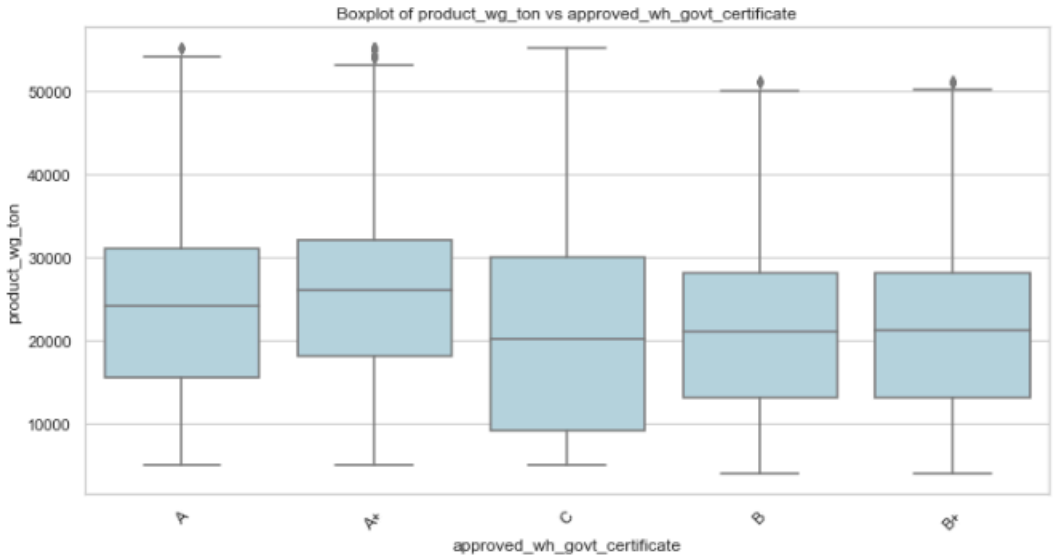
- **Location type vs Product weight**



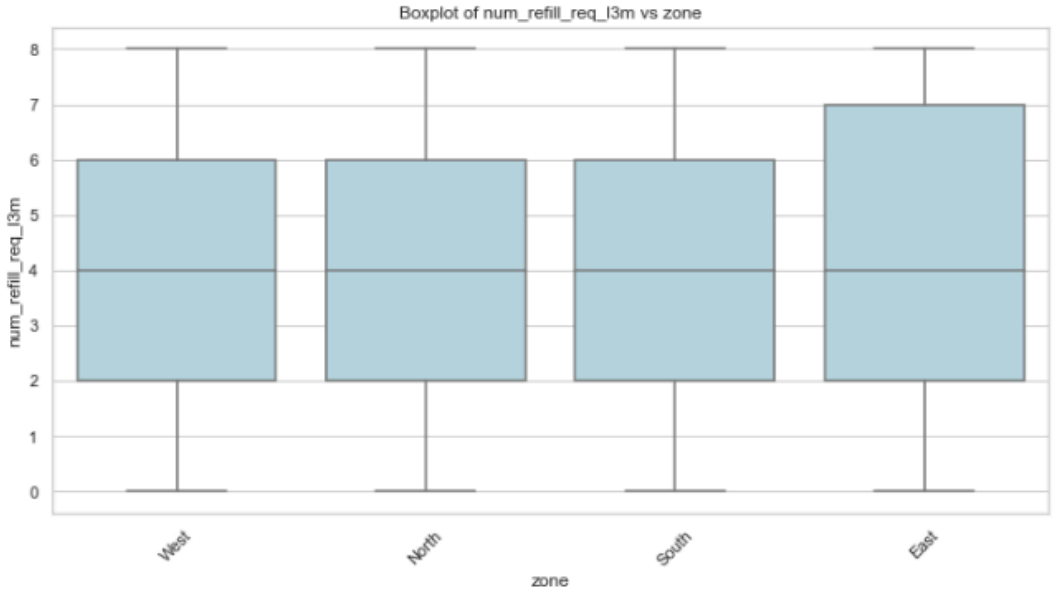
- **Existence of temperature regulation machine vs Product weight**



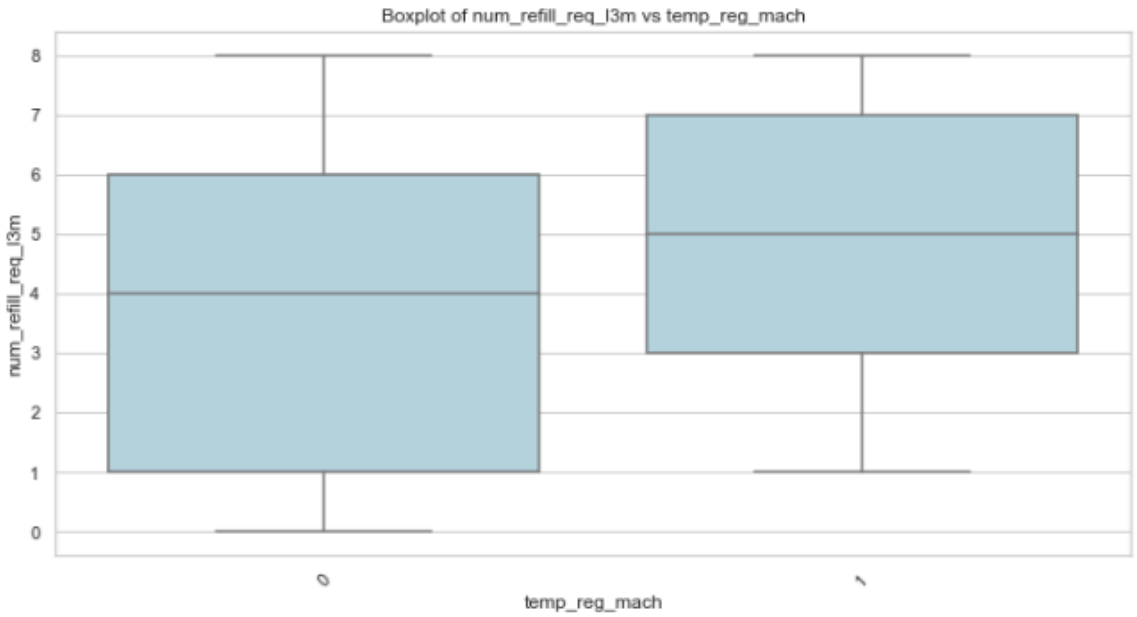
- **Government certification to the warehouse vs Product weight**



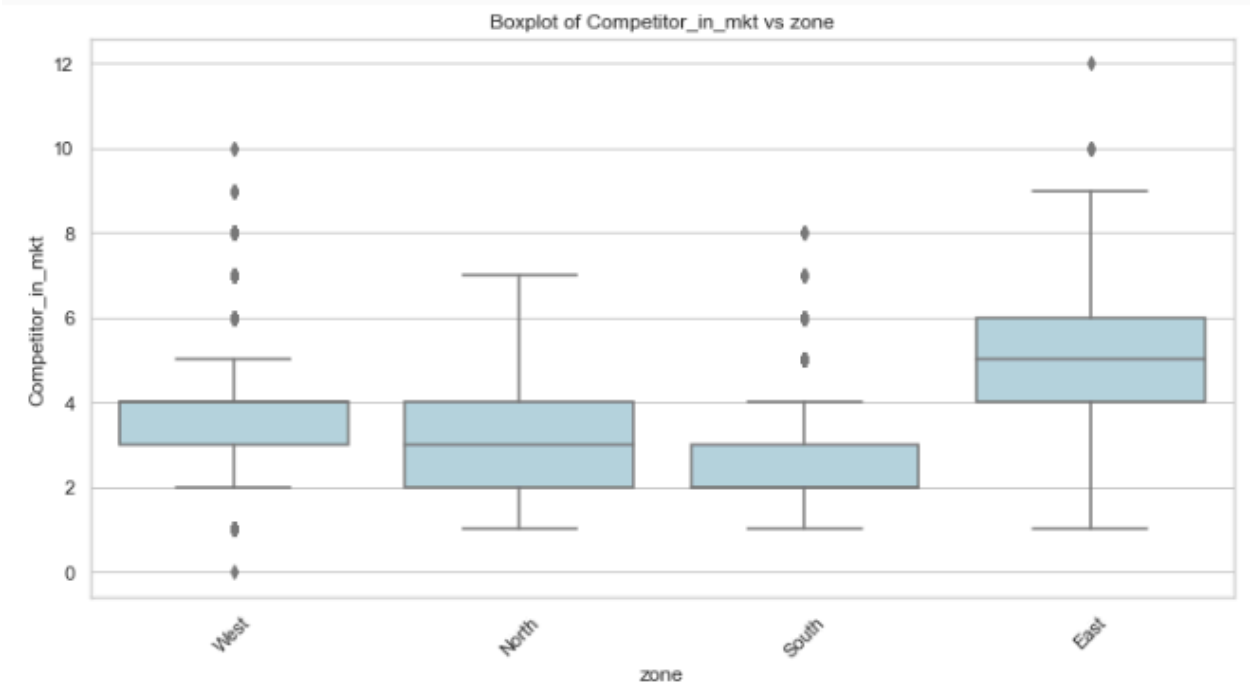
- **Number of refills required in last 3 months based on zone type**



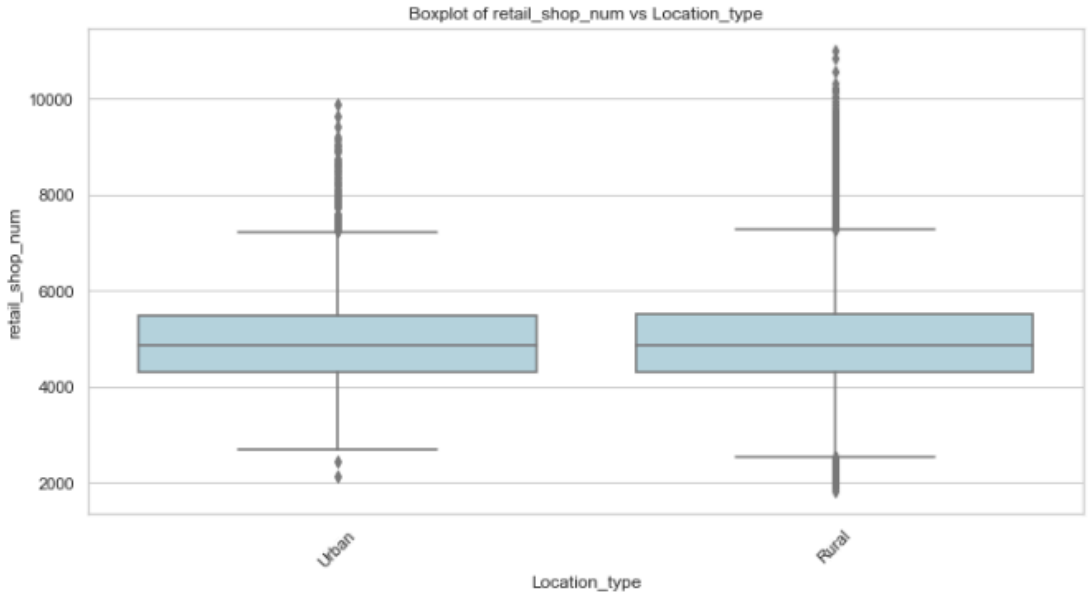
- **Number of refills required in last 3 months based on the availability of temperature regulating machines**



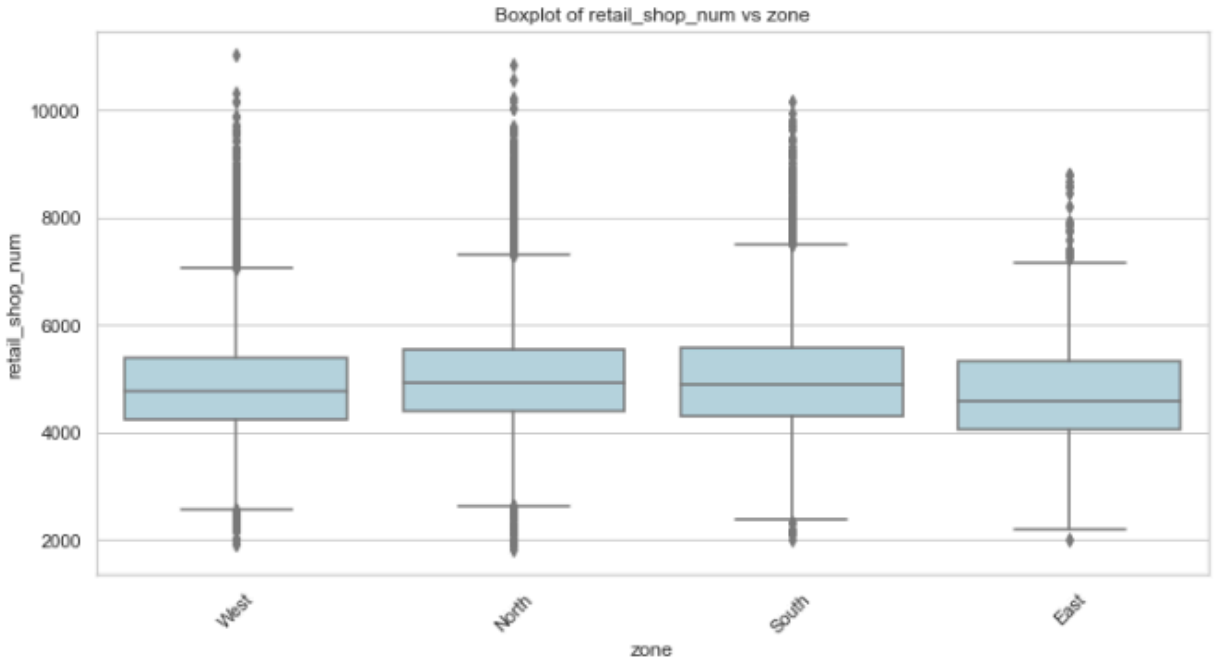
- **Number of competitors in the market based on zone type**



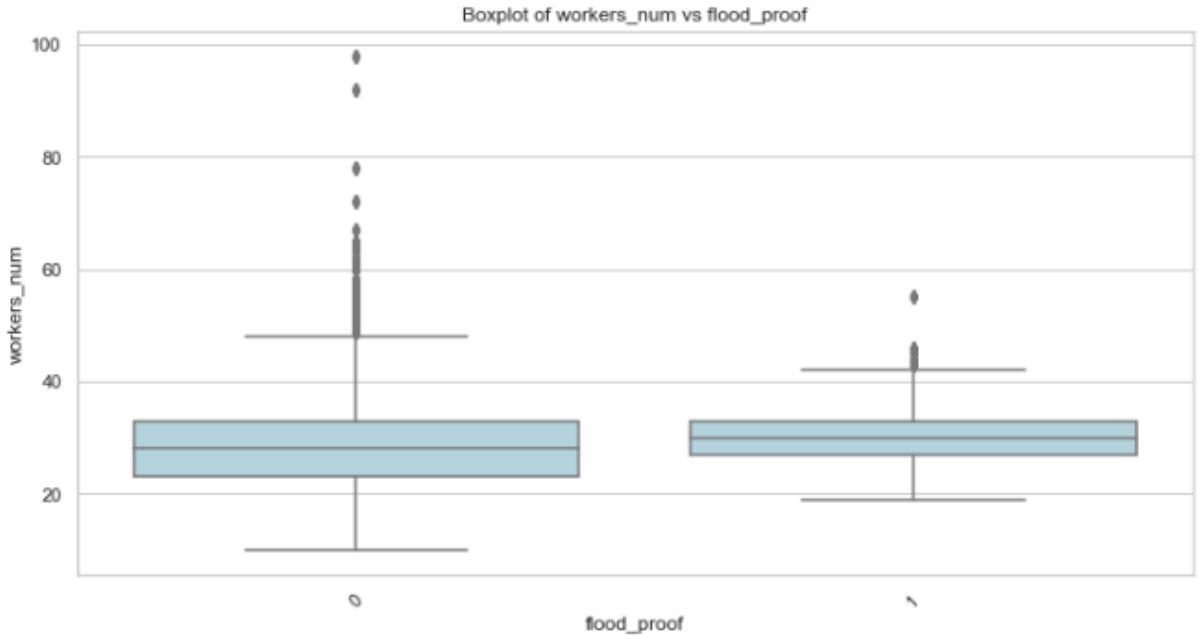
- **Number of retail shops that sell the product under the warehouse area based on location type**



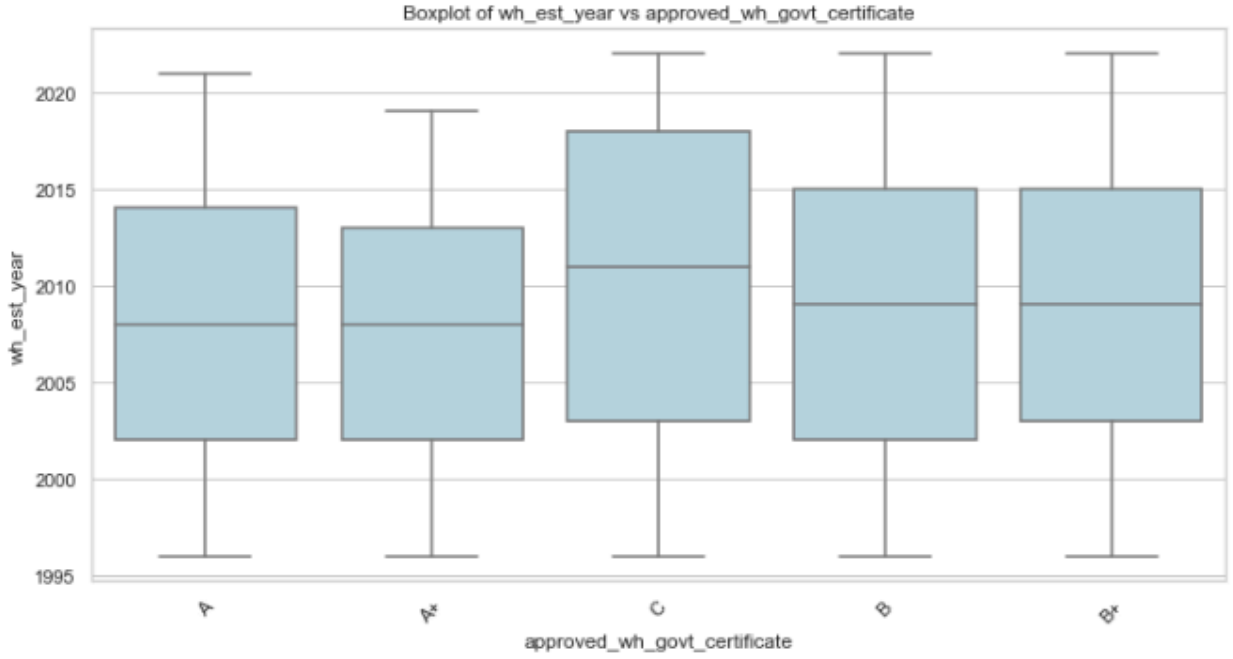
- **Number of retail shops that sell the product under the warehouse area based on zone type**



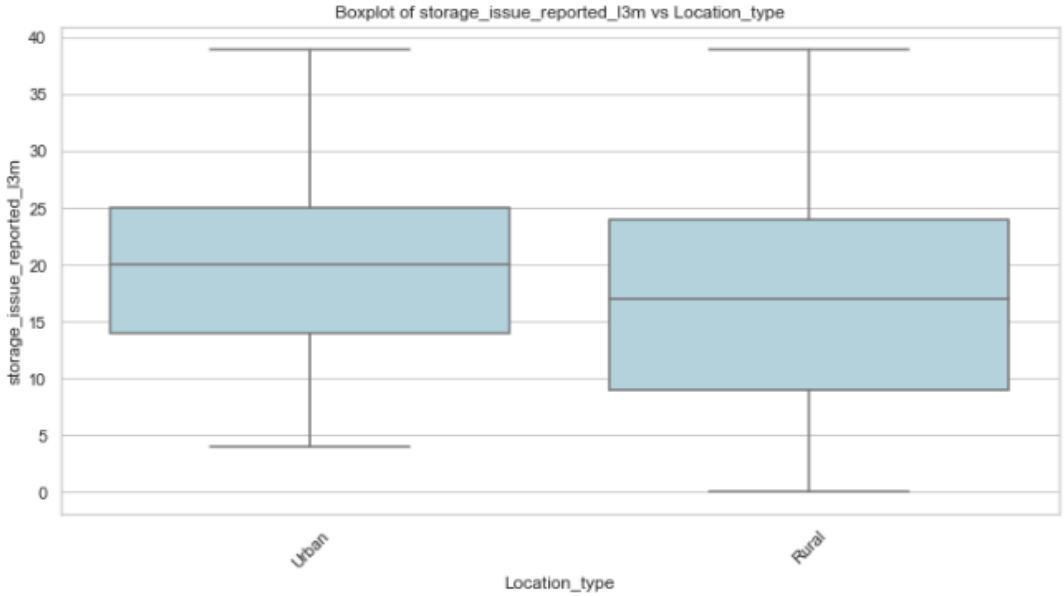
- **Number of workers working in the warehouse based of whether the warehouse was impacted by flood or not**



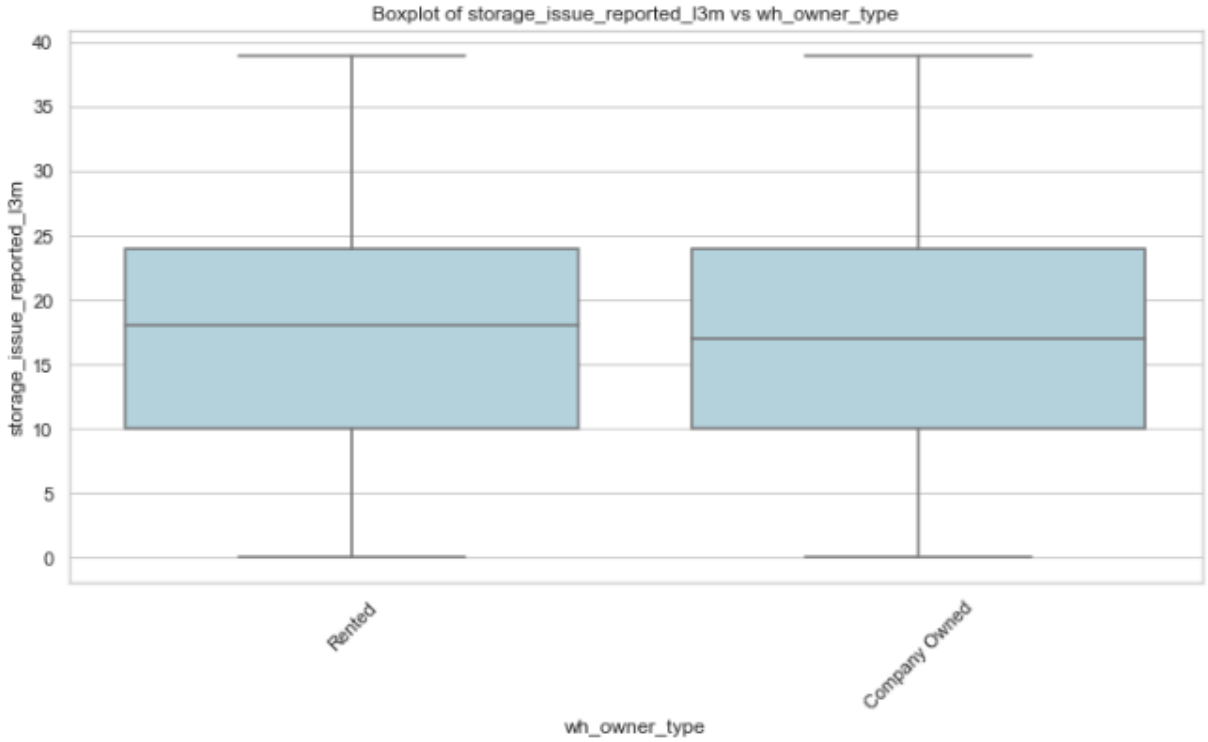
- **Government certification based on the establishment of the warehouse**



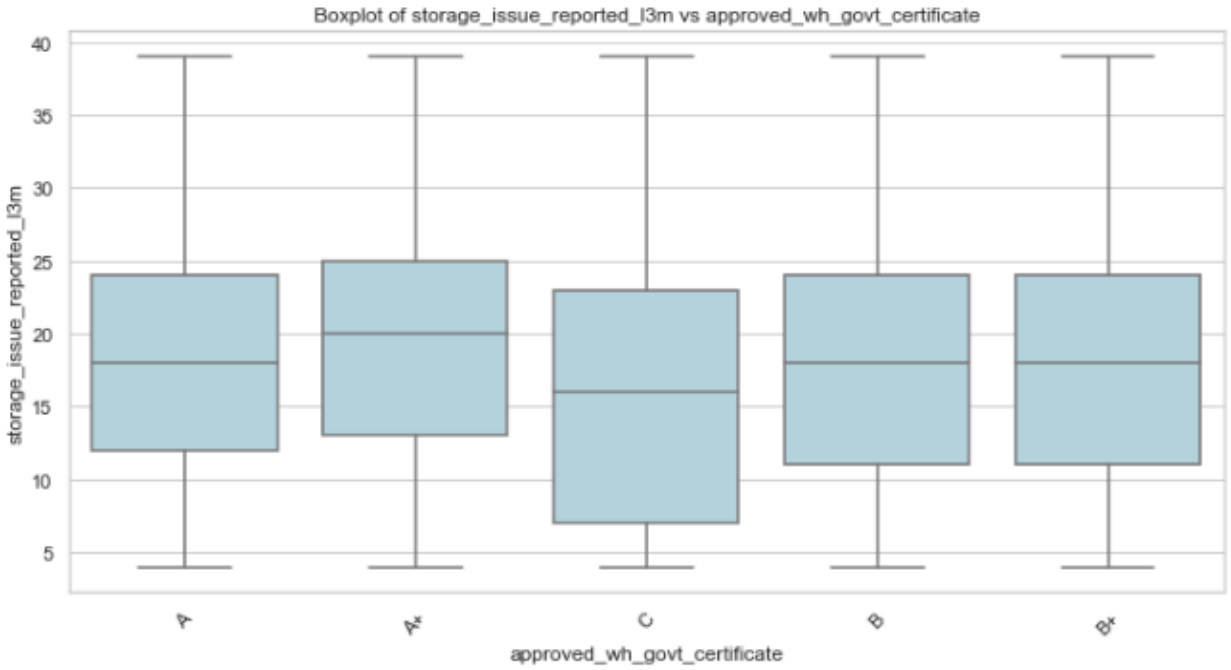
- **Number of reported storage issues in the last 3 months based on the location type**



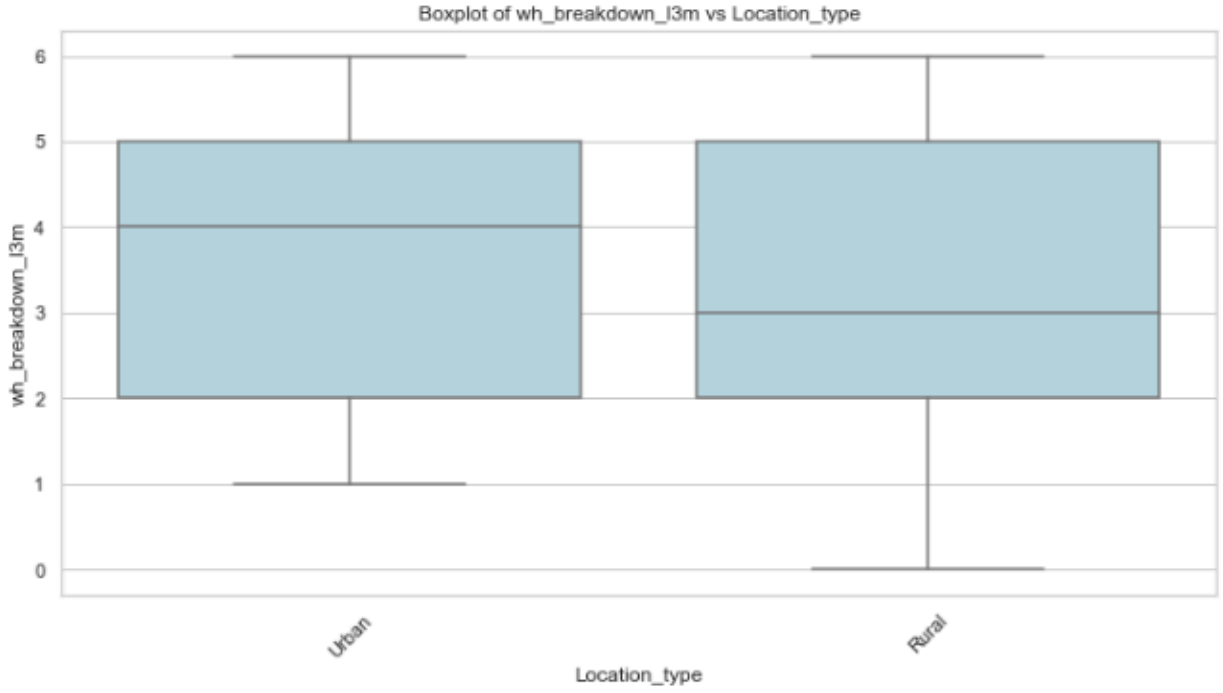
- **Number of reported storage issues in the last 3 months based on ownership of the warehouse**



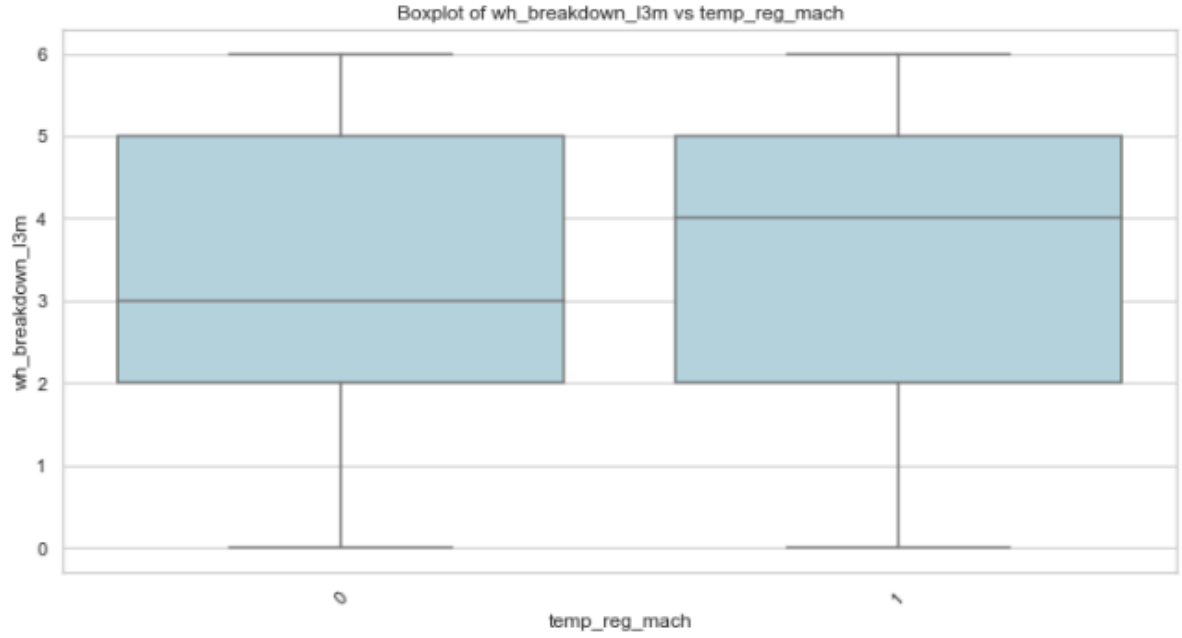
- **Number of reported storage issues in the last 3 months based on the government certification**



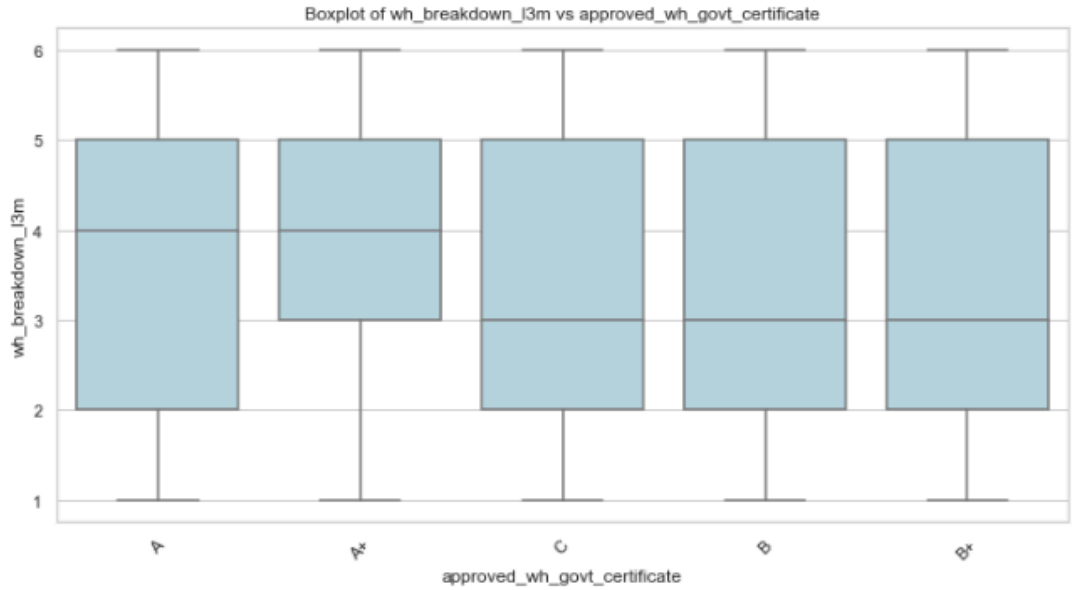
- **Number of times the warehouse faced a breakdown in the last 3 months based on location type**



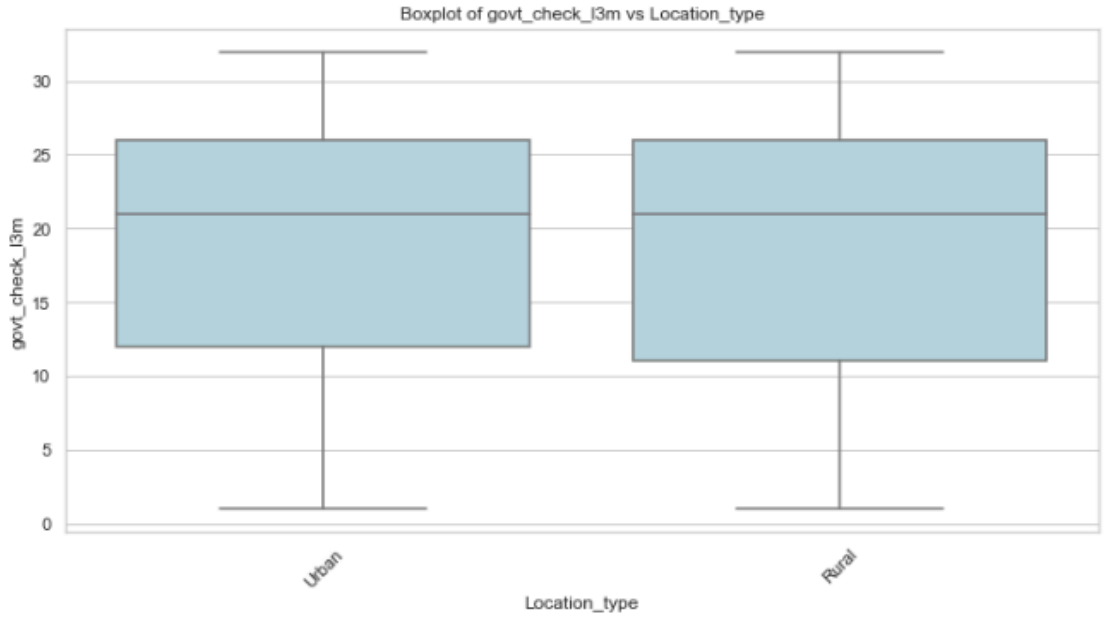
- **Number of times the warehouse faced a breakdown in the last 3 months based on the availability of the temperature regulating machine**



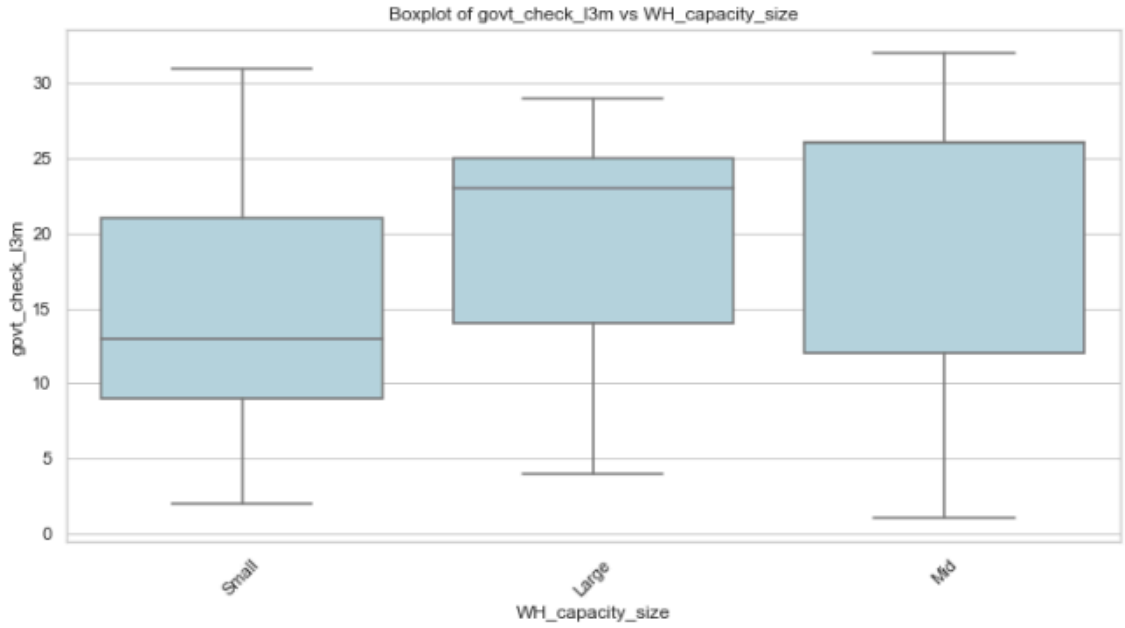
- **Number of times the warehouse faced a breakdown in the last 3 months based on the government certification**



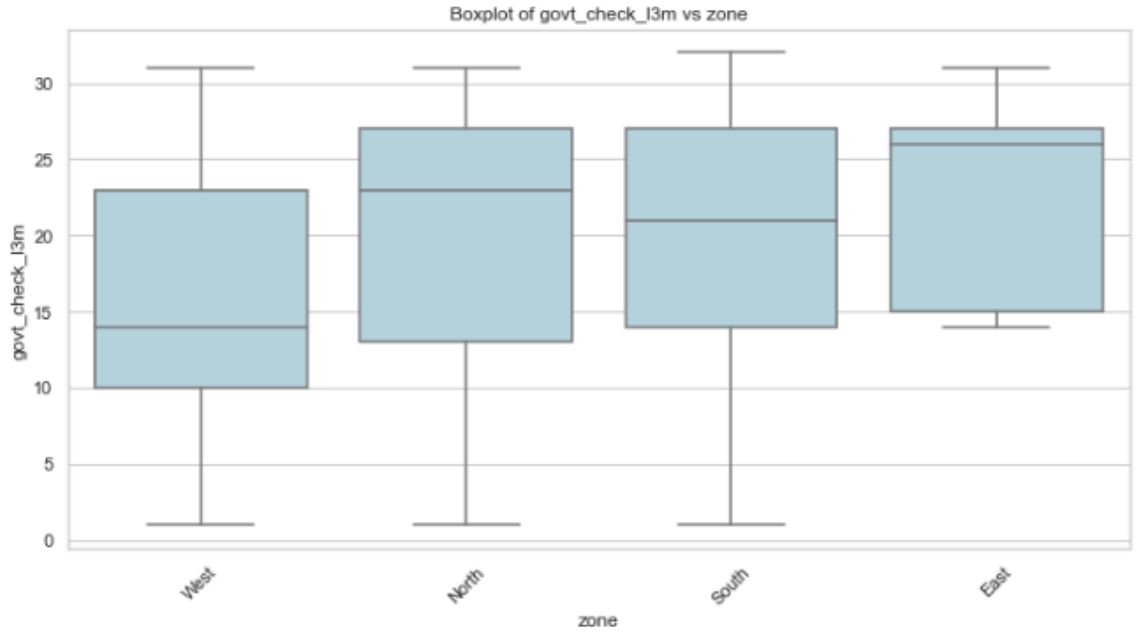
- **Number of times a government person checked a warehouse in the last 3 months based on the location type**



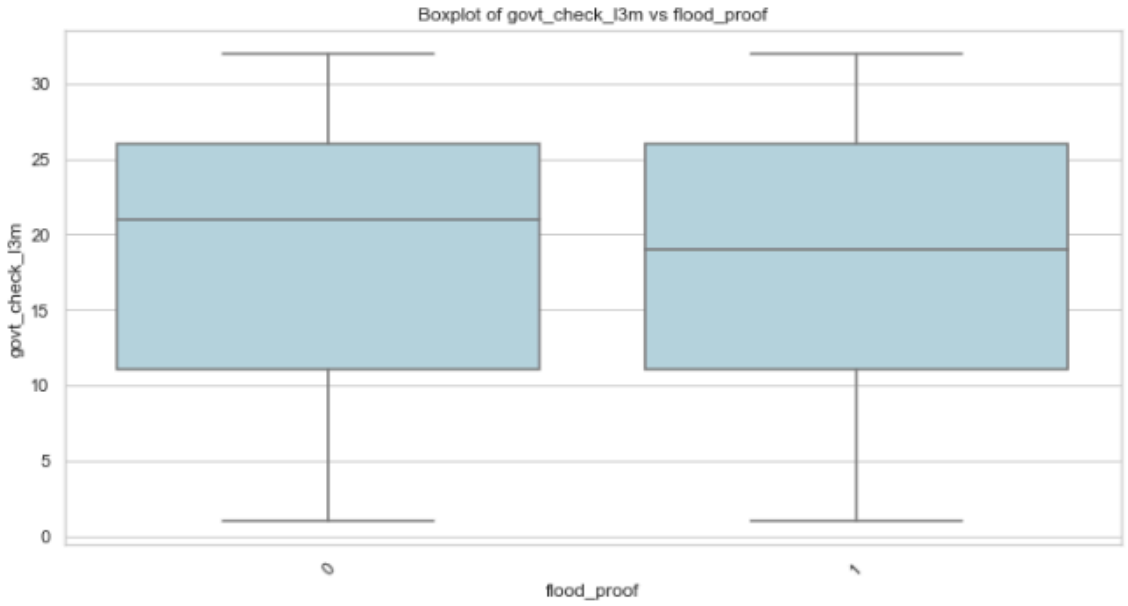
- **Number of times a government person checked a warehouse in the last 3 months based on the warehouse capacity size**



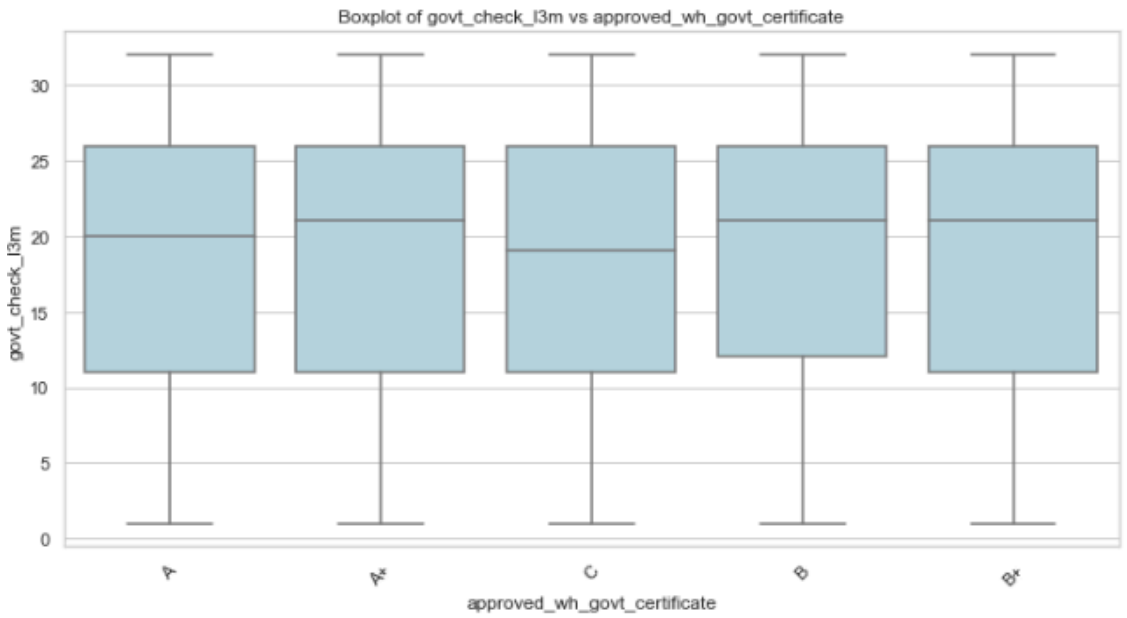
- **Number of times a government person checked a warehouse in the last 3 months based on the zone type**



- **Number of times a government person checked a warehouse in the last 3 months based on flood-proofing**

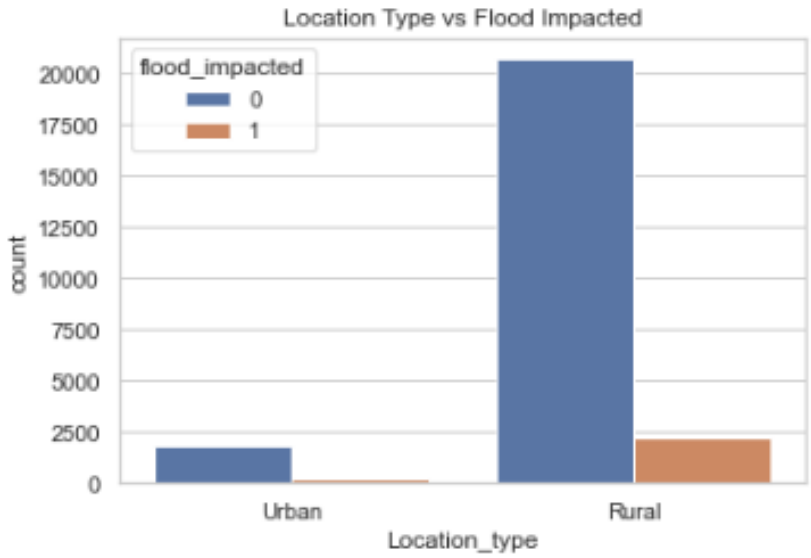


- **Number of times a government person checked a warehouse in the last 3 months based on government certification**

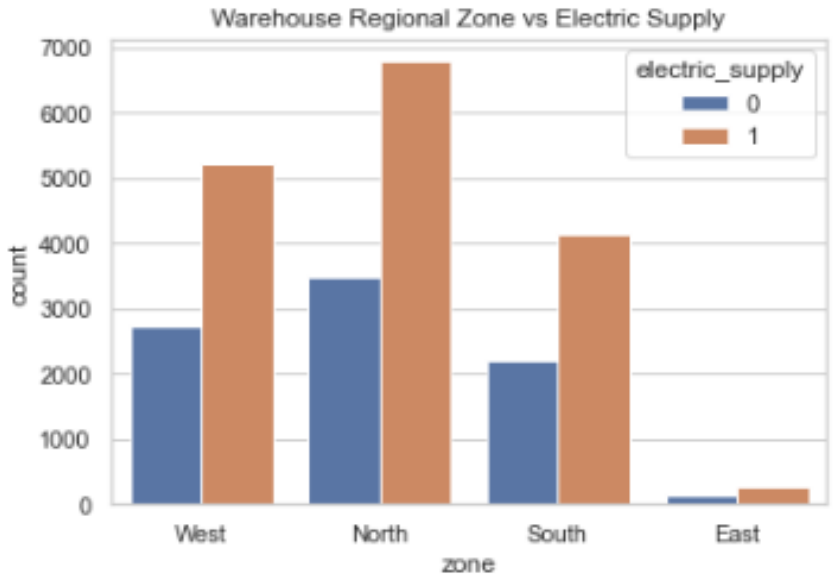


b) Multivariate analysis

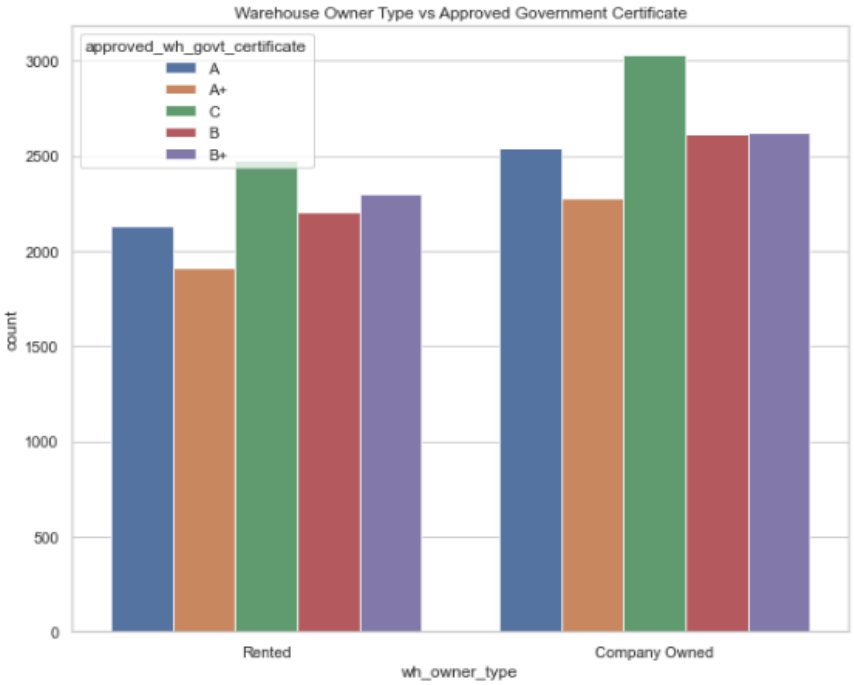
- **Impact of flood based on location type**



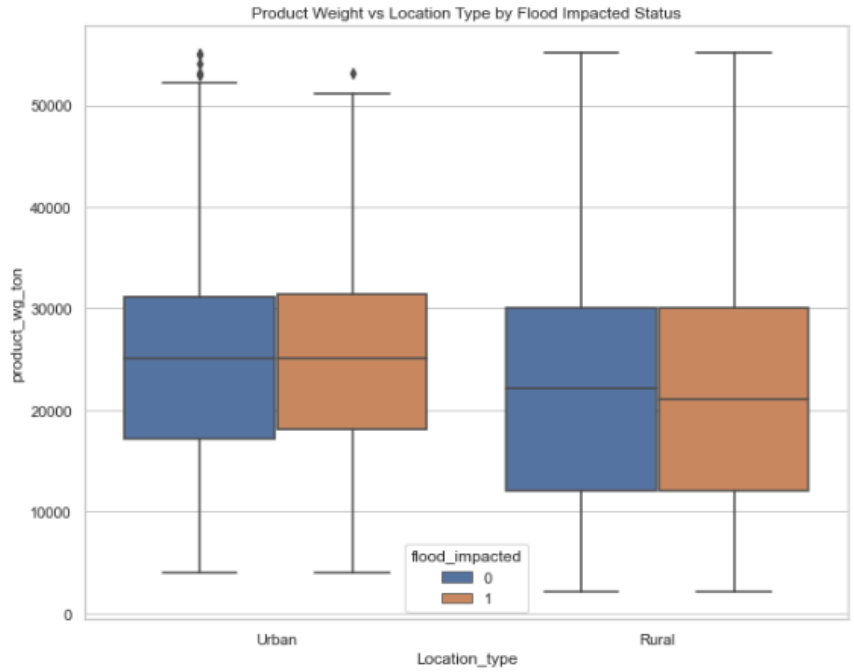
- **Electricity supply in warehouse based on zone type**



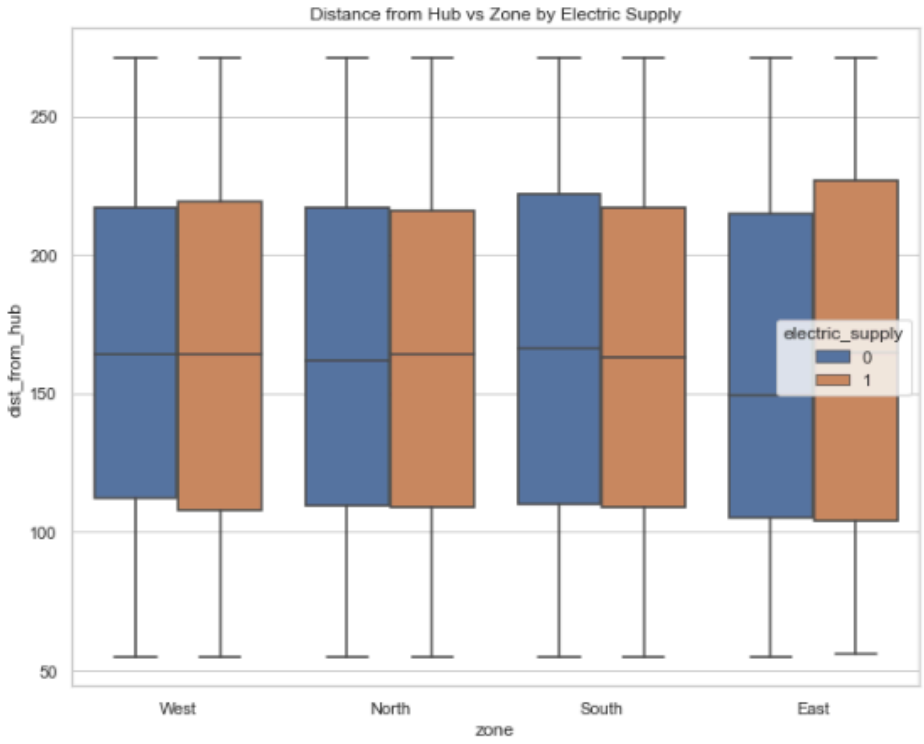
- **Government certification based on warehouse ownership**



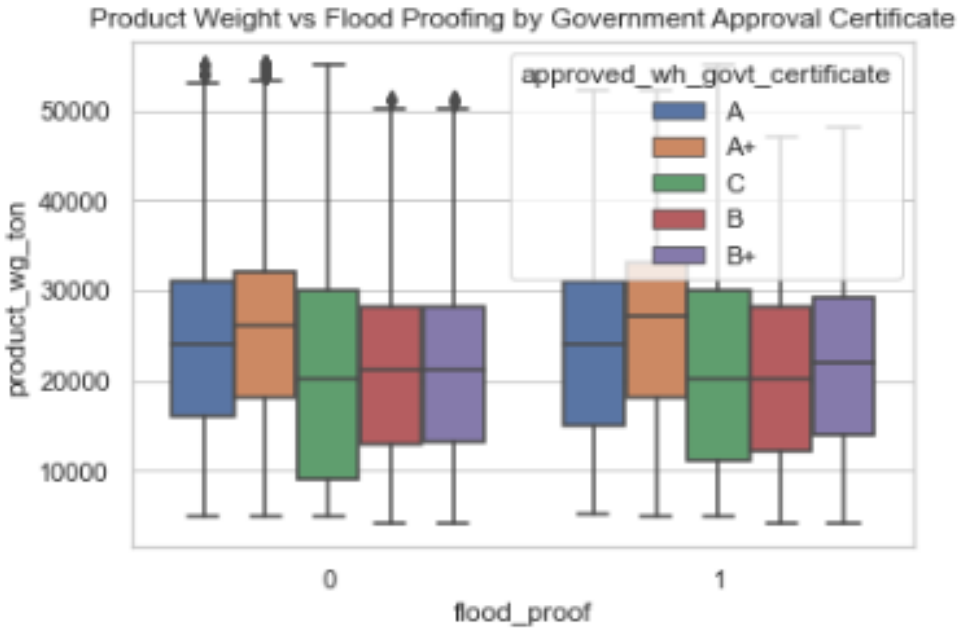
- **Relationship between location type, impact of flood, and product weight shipped to the warehouse**



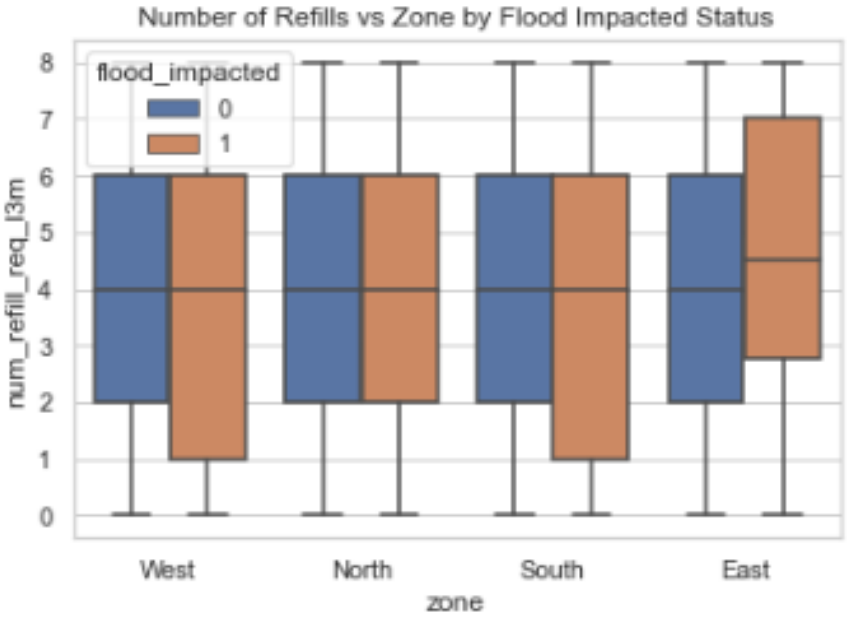
- **Relationship between electricity supply, zone type, and distance of warehouse from the hub**



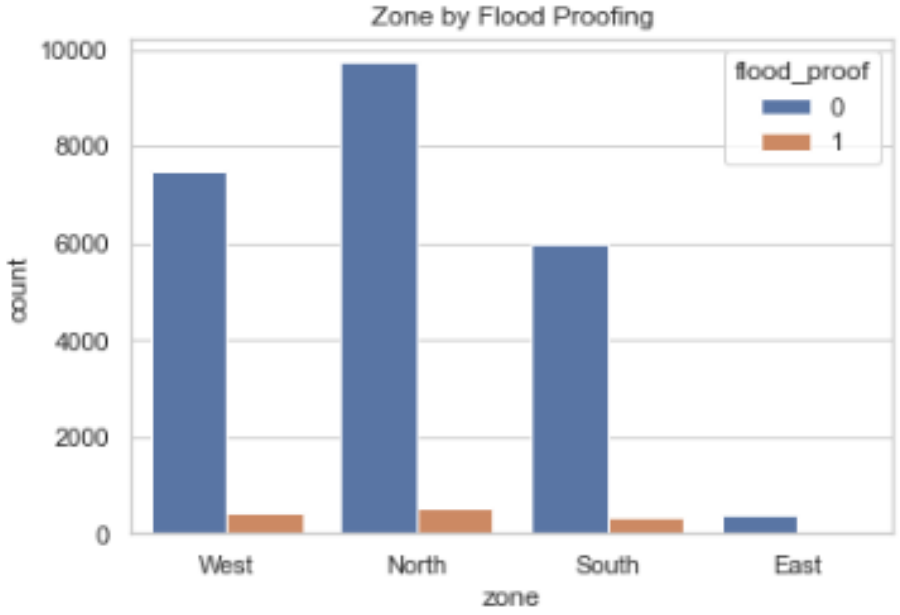
- **Relationship between floodproofing, government certification, and product weight shipped to the warehouse**



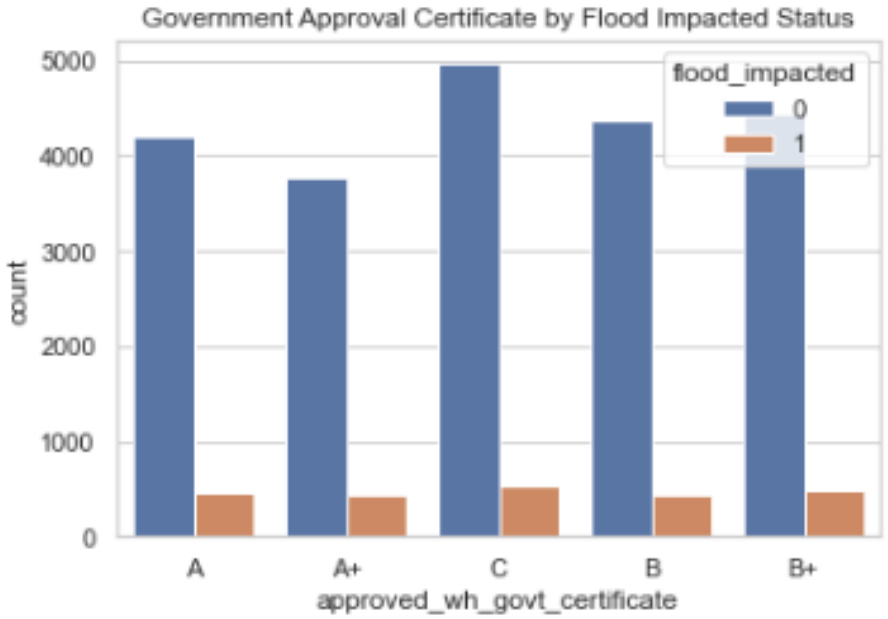
- **Relationship between impact of flood, zone type, and number of refills required in the last 3 months**



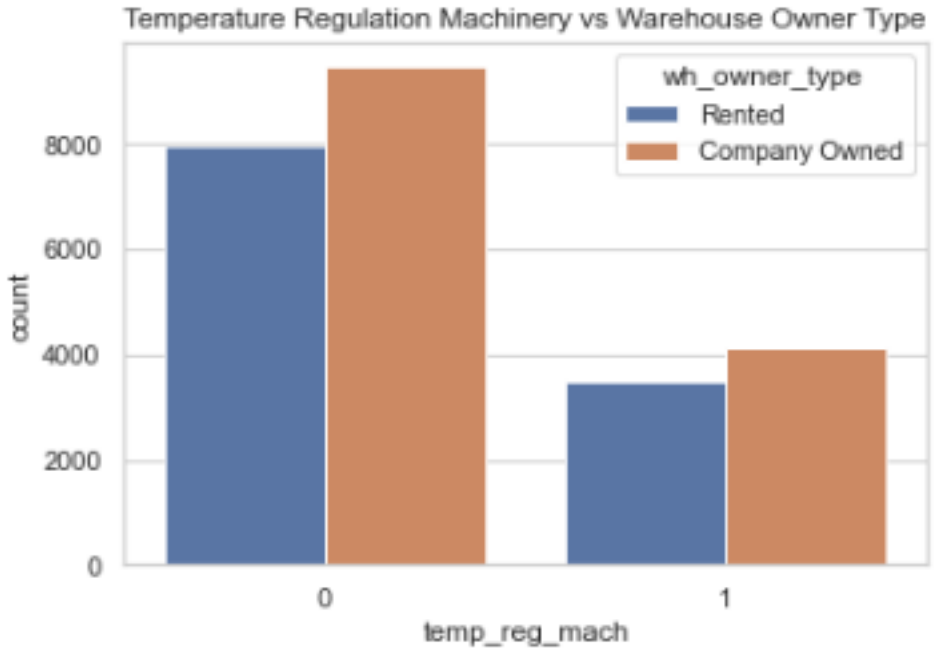
- **Relationship between floodproofing zone-wise**



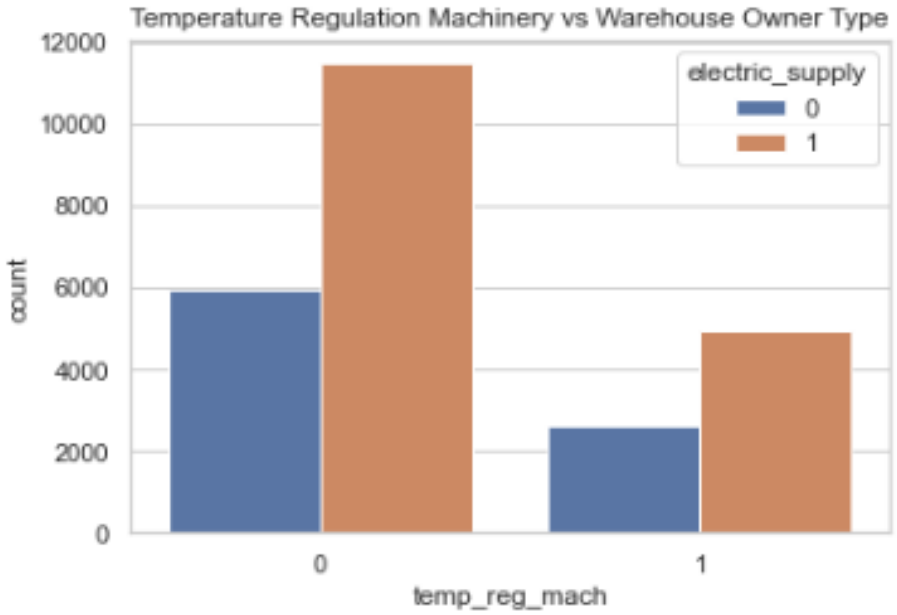
- **Relationship between government certification based on flood impact**



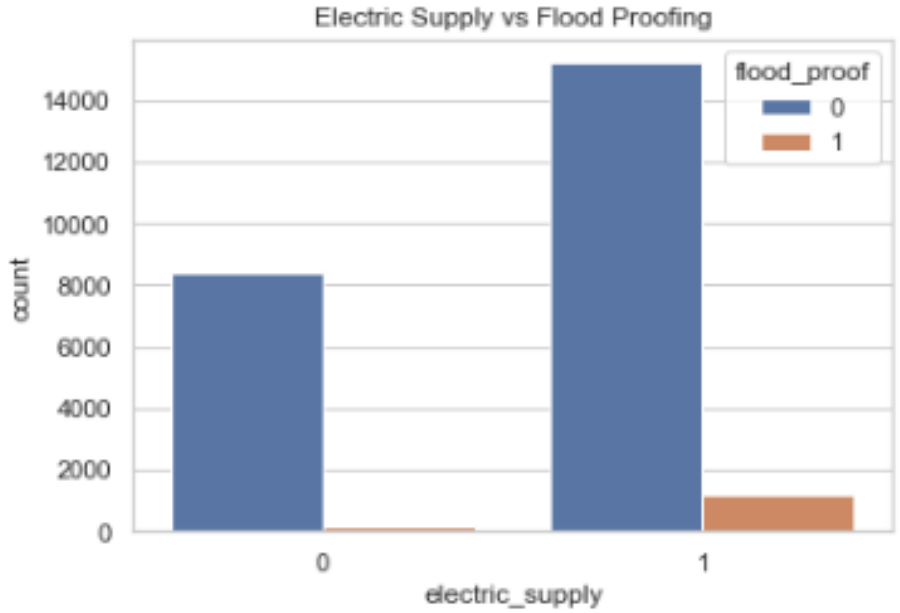
- **Availability of temperature-regulating machines based on ownership of the warehouse**



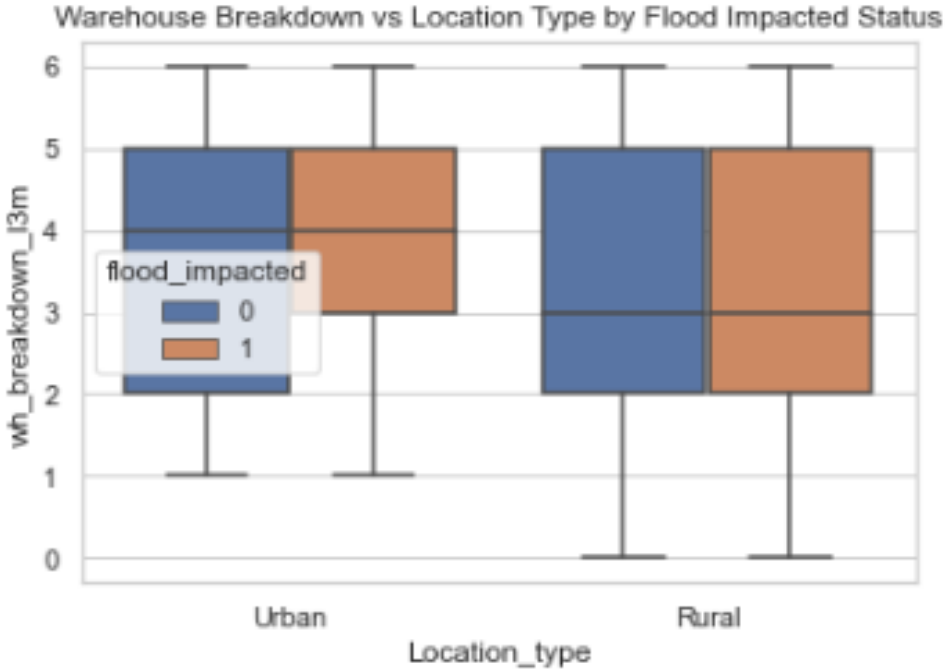
- **Availability of temperature-regulating machines based on electricity supply**



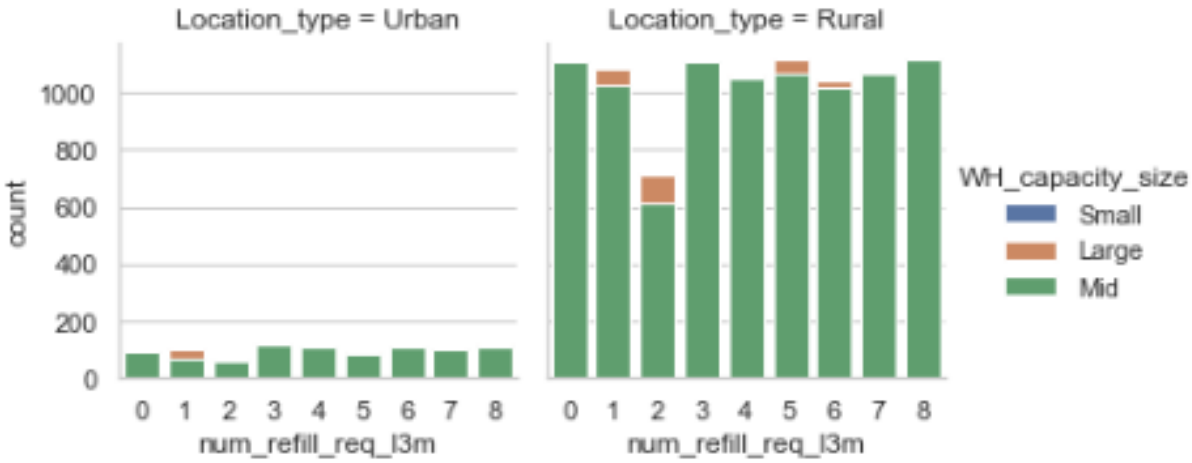
- **Availability of electricity supply based on floodproofing of warehouses**



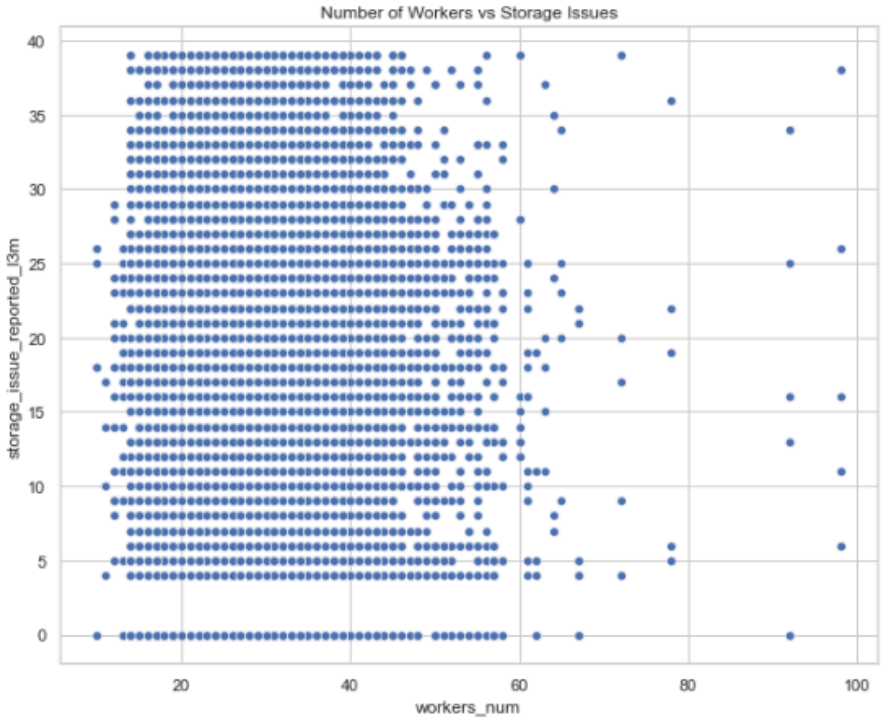
- **Relationship between location type, flood impact, and warehouse breakdown in last 3 months**



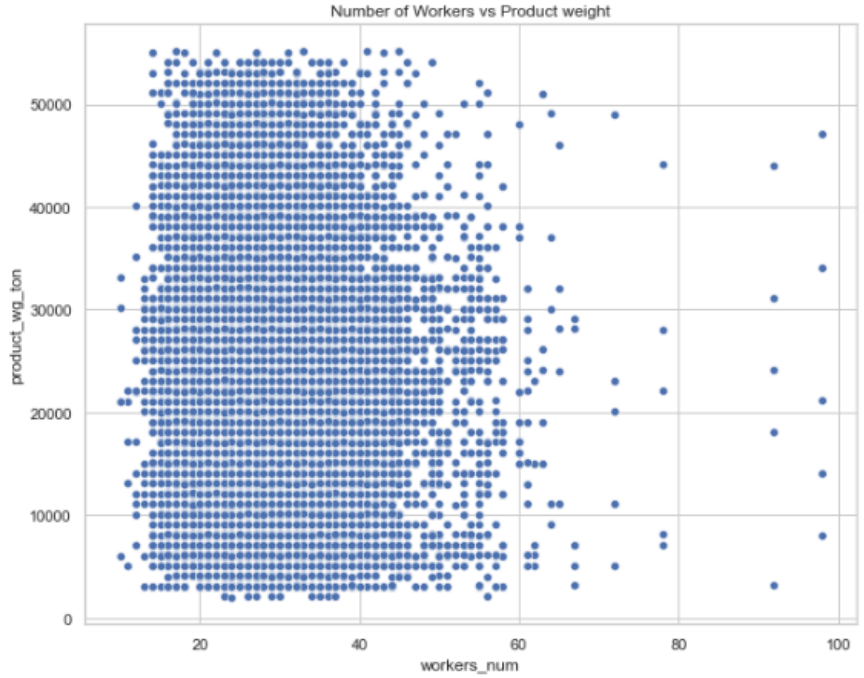
- **Number of refills required in last 3 months Vs Warehouse capacity size Vs Location type**



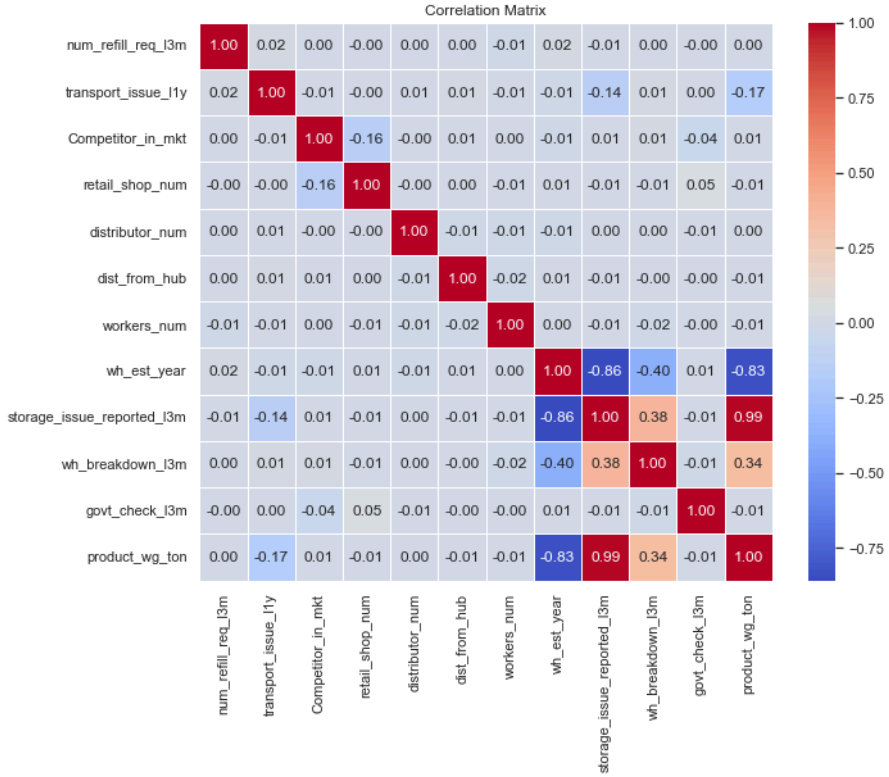
- **Correlation between the number of workers and storage issues reported in last months**



- **Correlation between number of workers and product weight shipped to warehouse**



- **Correlation matrix**



Data Cleaning and Preprocessing

Removal of unwanted variables

We dropped 'Ware_house_ID'and 'WH_Manager_ID' before doing the EDA, as IDs are mere labels.

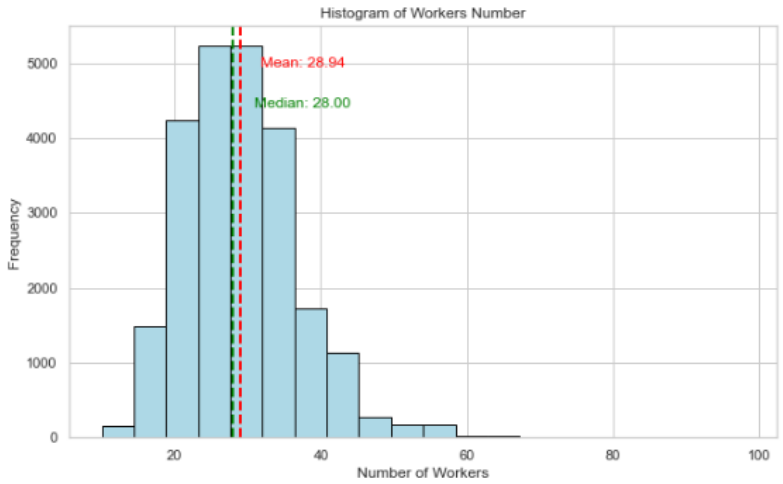
We also drop 'WH_regional_zone' as it is a subdivision of a zone, and it is unlikely that a zone's subdivision will impact our analysis.

Missing value treatment

There were missing values in the following columns:

Number of workers: 990 missing values
Warehouse established year: 11881 missing values
Government certification approval: 908 missing values

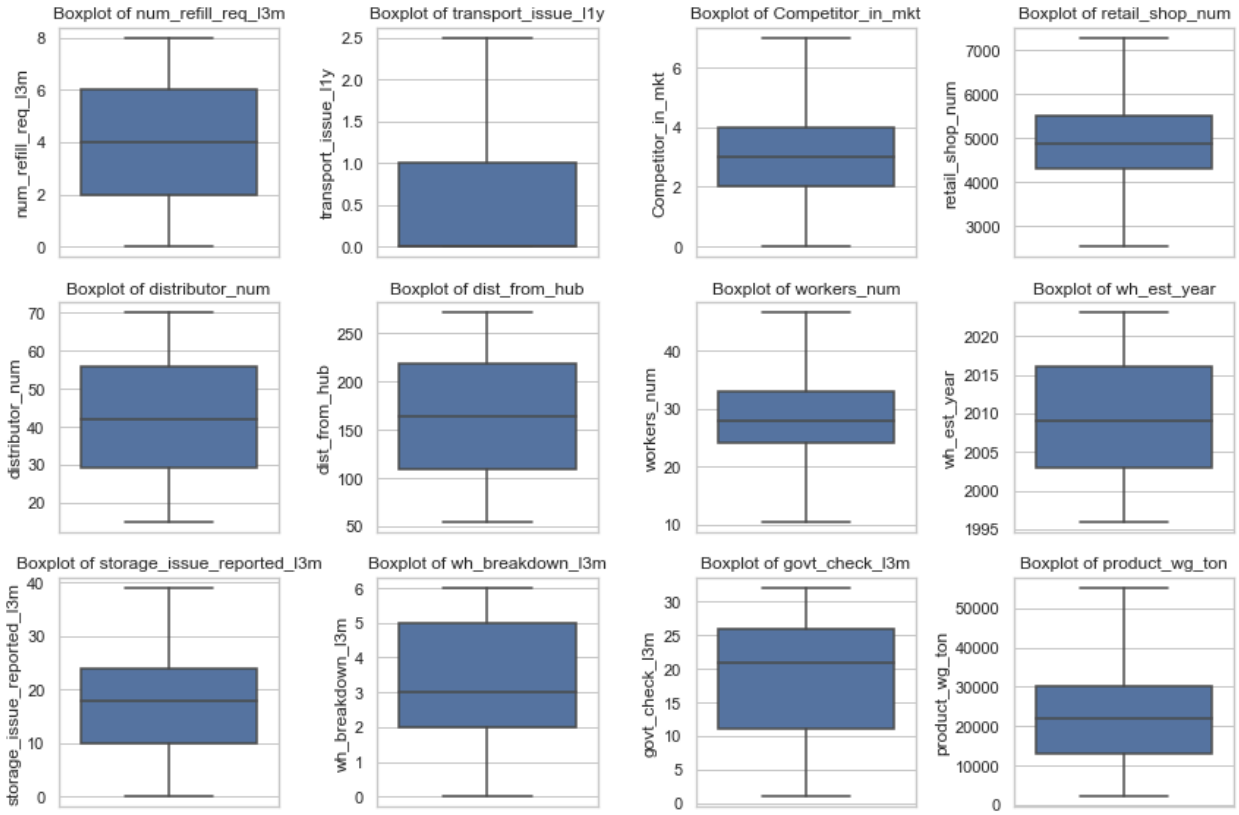
- We impute the missing values in wh_est_year using the KNN imputer.
    - This method finds the k-nearest neighbours in the dataset (based on other features) and imputes the missing value based on the mean or median of the nearest neighbours' `wh_est_year` values.
    - Ideally, this column should have been dropped since 47.5% of the data is missing. We choose not to drop it since we need to use this variable to add another variable called the warehouse_age which could be significant concerning warehouse-related issues.
- We impute the missing values of the 'workers_num' column with the median value since the data is skewed.



- Since 'approved_wh_govt_certificate' is a categorical variable, we impute the missing values of this column with the mode value.

Outlier treatment

We treat the outliers using the Inter Quartile Range method where we cap the values to the upper and lower boundaries. We can see that the outliers have been capped.



Variable transformation

We have converted the variables 'flood_impacted', 'flood_proof', 'electric_supply', and 'temp_reg_mach' from the integer type to (object) categorical type.

Addition of new variables

A new variable called warehouse_age is created using the variable wh_est_year, i.e., by subtracting the wh_est_year from the present date.
This is done because the age of the warehouse could play a significant role in warehouse-related issues.

Significance test

Significance test confirms that the observed correlation is almost certainly not due to random chance.

We use the Pearson correlation coefficient and p-value to check the significance. Pearson's Correlation Coefficient (r) measures the strength and direction of the linear relationship between two continuous variables.

- **Significance test between the newly added variable warehouse_age and the target variable product_wg_ton**

  Pearson correlation coefficient: 0.6050
  p-value: 0.0000 (the statistical software or library could have rounded the p-value to 0 due to it being extremely small.)

  A Pearson correlation coefficient of 0.6050 indicates a moderate positive linear relationship between two variables.

  Since p value is 0, the relationship between 'warehouse_age' and 'product_wg_ton' is significant.

- **Significance test between warehouse_age and wh_breakdown_l3m**

  Pearson correlation coefficient: 0.2882
  p-value: 0.0000

  The correlation coefficient of 0.2882 suggests a weak positive linear relationship between the two variables. While the relationship between the two variables is weak, it is real and not due to random chance.

  The p-value of 0.0000 (or a very small value close to zero) indicates that the correlation is statistically significant. In other words, the likelihood that this correlation occurred by random chance is extremely low.

ANOVA test to check the significance between categorical variables and the target variable.

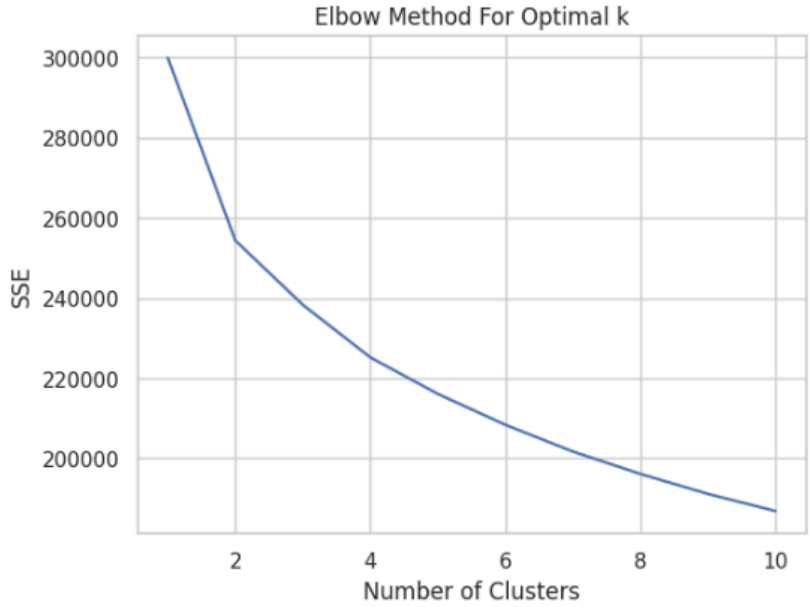
The F-statistic compares the variance between the groups to the variance within the groups. A larger F-statistic indicates a greater difference between the groups relative to within the groups.

A low p-value (typically < 0.05) suggests that there is a statistically significant difference in the means of `product_wg_ton` across the categories of the variable.

- The variable 'flood_impacted' is not significant (p-value: 0.7162).
- The variable 'flood_proof' is not significant (p-value: 0.9444).
- The variable 'electric_supply' is not significant (p-value: 0.7327).
- The variable 'temp_reg_mach' is significant (p-value: 0.0000).
- The variable 'approved_wh_govt_certificate' is significant (p-value: 0.0000).

Clustering of warehouse based on demand

- Prior to clustering, we scale the data.
- We use the k-means clustering method to group the data.
- From the elbow graph we can observe that 3 is an optimal number of clusters.



- The clustering of the data frame based on the numerical values is as follows:
    - Cluster 2 has an average warehouse age of 18, the older warehouses belong to this cluster and Cluster 1 has an average warehouse age of 10, the relatively newer warehouses belong to this cluster.
    - Intuitively, we can see that Cluster 2 has a higer average number of warehouse breakdowns and Cluster 1 has a lower average number of warehouse breakdowns.
    - Similarly, another intuitive observation is that Cluster 2 has more storage issues reported in last 3 months and Cluster 1 has a relatively low storage issues reported in last 3 months.

- The average weight of the product (noodles) shipped to warehouses belonging to Cluster 2 is the highest, followed by Cluster 0, and then Cluster 1.

Business Insights from EDA

Data imbalance

- Frequency distribution of location type: The majority of the data (22957) is from the rural areas and only 2043 are from the urban areas.
- The data from the East zone is merely 429 and from the North, it is a whopping 10278. There is a huge gap between the numbers.
- Regarding data on transport issues: Most data, almost half, on transport issues in the last year have been ranked 0, implying there were no issues.
- Of all the data on warehouses, only 2454 of them have been impacted by flood, remaining 22546 have not been impacted by floods.
- Only 1366 have been floodproofed, and the remaining 23634 are not.
- 16422 warehouses have electricity supply and 8578 don't.
- 17418 warehouses do not have temperature-regulating machines and only 7582 have.

Such extremities and imbalances could lead to biased insights. Models may overly generalize based on the majority class (rural and North zone), potentially ignoring patterns in the minority classes (urban and East zone).

With a vast majority of warehouses not impacted by floods and not floodproofed, any analysis predicting flood impact or recommending floodproofing may struggle to learn from the limited data in the minority class.

The skewed distribution, like most transport issues being ranked 0 or the majority of warehouses having electricity but lacking temperature-regulating machines, can lead to skewed insights. If the analysis is not adjusted, it might lead to underreporting or overemphasizing specific issues.

For instance, a model predicting warehouse flood impact might just predict "no impact" to get higher accuracy but fail to correctly predict the rare but crucial cases of impact.

This could also lead to incorrect decisions, like assuming most warehouses are in rural areas or have no transport issues when there might be significant, albeit rare, exceptions.

Mitigating the Impact:

- **Resampling Techniques:** Employ oversampling or undersampling to balance the data.

- **Weight Adjustments:** Adjust weights in the model to give more importance to minority classes.
- **Anomaly Detection:** Use specialized techniques to detect patterns in minority classes.

Balancing the data or using techniques that account for the imbalance can help ensure our analysis is robust and reliable.

Other observations:

- The median product weight that is shipped to warehouses is higher for urban locations than for rural locations.
- The median product weight that is shipped to warehouses is higher for those with temperature-regulating machines than those without.
- The government certification for warehouses is the best, i.e. A+, for those where the median product weight shipped is higher and the lower, i.e. C, for those where the median product weight is less.
- The frequency distribution of government certification of C is the highest and A+ is the lowest in numbers.
- The highest frequency of warehouse breakdown in the last 3 months is 2 to 3 times.
- The East zone has the maximum frequency of refills required in the last 3 months.
- The median number of refills and 50% of the frequency required in the last 3 months is higher for warehouses with temperature-regulating machines
- The East zone has a higher median number of competitors in the market.
- The urban area has a higher median number of storage issues reported in the last 3 months
- The urban locations have a higher median warehouse breakdown in last 3 months' value and a wider range of values. The rural locations have a lower median warehouse breakdown in the last 3 months' value and a narrower range of values.
- The median value of `warehouse breakdown in the last 3 months` appears to be higher in the group with the approved certificate (A+) compared to the group with C certification
- The median number of government checks in the last 3 months is higher for medium and large warehouses.
- The East zone has the highest median number of government checks in the last 3 months followed by the North zone.
- The rural locations have a relatively higher flood impact than the urban locations.
- The availability of electricity supply is the highest in the North zone and least in the East zone.
- The A certification by the government is higher for company-owned warehouses than rented ones, but so is the C certification.
- There seems to be a relationship between the number of refills required in the last 3 months based on the impact of flood, and it is significantly higher for the East zone.
- There is a strong positive correlation between storage issues reported in the last 3 months and product weight.
- There is a strong negative correlation between storage issues reported in the last 3 months and warehouse established year.

- There is a strong negative correlation between warehouse established year and product weight.

Business insights using clustering

1. **Aging Infrastructure Challenges (Cluster 2)**:
    - The older warehouses in Cluster 2, with an average age of 18 years, appear to be facing significant operational challenges, including a higher frequency of breakdowns and storage issues. This indicates that the infrastructure in these warehouses might be deteriorating, leading to higher maintenance costs and potential disruptions in operations.
    - **Actionable Insight**: Consider investing in the renovation or replacement of aging warehouses in Cluster 2 to improve reliability and reduce operational risks. Additionally, implementing preventive maintenance programs could help mitigate the frequency of breakdowns.
2. **Efficiency of Newer Warehouses (Cluster 1)**:
    - The relatively newer warehouses in Cluster 1, with an average age of 10 years, exhibit fewer breakdowns and storage issues, indicating that they are more efficient and reliable in their operations.
    - **Actionable Insight**: The practices and technologies used in Cluster 1 could be analyzed and potentially replicated or adapted in older warehouses (Cluster 2) to improve overall efficiency.
3. **High Product Weight Distribution in Older Warehouses (Cluster 2)**:
    - Despite the operational challenges, Cluster 2 warehouses are handling the highest average weight of products shipped, suggesting that these warehouses might be critical hubs in the distribution network.
    - **Actionable Insight**: The high volume of product distribution in Cluster 2 warrants careful monitoring. Ensuring that these warehouses are kept in optimal condition is essential to avoid disruptions in the supply chain, especially given their pivotal role in product distribution.
4. **Potential Overloading or Mismanagement**:
    - The combination of older age, more frequent breakdowns, higher storage issues, and the highest average product weight suggests that Cluster 2 warehouses may be overburdened. The operational strain might be contributing to the increased incidence of problems.
    - **Actionable Insight**: Consider redistributing the load more evenly across different clusters or reducing the burden on Cluster 2 warehouses. Additionally, implementing more robust management practices or introducing automated systems could help mitigate some of these issues.

These insights suggest that a strategic focus on infrastructure management, operational efficiency, and load balancing across the network could improve the overall performance and reliability of the warehouse system.

Other business insights and recommendations

1. Regional Disparities

Insight: The North zone dominates in terms of warehouse data, while the East zone shows lower values across multiple metrics. The rural areas have a higher frequency of warehouse data compared to urban areas, but rural locations also face higher flood impact.

Recommendation: Focus on the North zone for expansion and resource allocation, given its higher warehouse count and electricity supply. For the East zone, address issues like low electricity availability and high number of refills required. Rural areas should be prioritized for flood risk management improvements.

2. Impact of Temperature-Regulating Machines

Insight: Warehouses with temperature-regulating machines exhibit higher median product weights and require more frequent refills. These warehouses also have higher median storage issues but are generally better certified by the government.

Recommendation: Invest in temperature-regulating machines to handle higher product weights and improve government certification ratings. However, monitor and manage the associated increase in storage issues and refills to optimize operational efficiency.

3. Government Certification and Warehouse Performance

Insight: Warehouses with higher median product weights tend to have better government certifications (A+), while those with lower product weights often have lower certifications (C). Additionally, warehouses with A+ certifications experience higher median breakdowns.

Recommendation: Ensure that warehouses are prepared to handle the higher product weights that come with A+ certification. Evaluate whether investing in A+ certification is beneficial despite the potential for higher breakdown rates.

4. Urban vs. Rural Performance

Insight: Urban warehouses report more storage issues and higher breakdown rates compared to rural ones. However, urban warehouses also have a higher median product weight shipped.

Recommendation: Implement strategies to reduce breakdowns and storage issues in urban warehouses. This might include improved infrastructure, better training, or enhanced maintenance protocols.

5. Correlation Insights

- **Product Weight and Storage Issues:** There is a strong positive correlation, indicating that as product weight increases, storage issues also rise. This suggests that heavier products require more careful handling and storage strategies.
- **Warehouse Age and Product Weight:** There is a strong negative correlation, meaning older warehouses may not handle heavier products as effectively. Consider upgrading or replacing older warehouses to better manage heavier products.
- **Warehouse Age and Storage Issues:** A strong negative correlation suggests that newer warehouses might face fewer storage issues. This could indicate that modern facilities have better design and technology integration.

Recommendation: Invest in upgrading older warehouses and ensure that new warehouses are designed to handle both high product weights and minimize storage issues.

6. Flood Impact and Refills

Insight: The East zone shows a significant correlation between flood impact and the number of refills required, suggesting that flood-affected warehouses face more operational challenges.

Recommendation: Implement flood mitigation measures in the East zone and establish contingency plans to manage increased refills and operational disruptions due to flooding.

7. Electricity Supply and Zone Performance

Insight: The North zone has the highest availability of electricity supply, which correlates with its dominant warehouse presence. The East zone, with the least electricity supply, faces challenges related to refills and operational efficiency.

Recommendation: Enhance electricity infrastructure in the East zone to improve warehouse performance and reduce the need for frequent refills. Consider investing in alternative energy sources or backup systems for critical operations.

8. Warehouse Size and Government Checks

Insight: Medium and large warehouses undergo more government checks compared to smaller ones. The East zone has the highest median number of checks, followed by the North zone.

Recommendation: Ensure that medium and large warehouses are compliant with regulatory requirements to pass government checks efficiently. Address any issues highlighted during these checks to avoid operational disruptions.

- Expand and focus on the North zone for growth opportunities.
- Invest in temperature-regulating technology, but manage its impact on storage issues.
- Prioritize flood risk management, especially in the East zone.
- Upgrade older warehouses and leverage the benefits of modern facilities.
- Enhance electricity infrastructure in the East zone.
- Ensure compliance and address issues identified during government checks.

Model Building and Model Validation

Data preparation for modelling and an overview

- We use one-hot encoding to encode the categorical variables.
- We split the data into train and test data in a 70:30 ratio.

Choice of models and the rationale for choosing

Linear regression

- **Simplicity and Interpretability**: It assumes a linear relationship between the dependent variable (e.g., product weight in tons) and independent variables (e.g., warehouse capacity, distance from hub, etc.). This model offers clear insights into how each feature impacts the target variable.
- **Ease of implementation**: It is quick to implement and serves as a benchmark for evaluating more complex models.

Ridge and Lasso regression

- **Handling Multicollinearity**: Supply chain data might contain correlated variables (e.g., number of workers and warehouse capacity). Ridge regression penalizes large coefficients, reducing overfitting due to multicollinearity.
- **Feature Selection (Lasso)**: Lasso regression introduces sparsity by shrinking some coefficients to zero, which helps identify key drivers in the supply chain performance while eliminating less impactful variables.
- **Balancing Bias-Variance Tradeoff**: These models help prevent overfitting through regularization, making them more robust on unseen data.

Decision trees

- **Non-linear Relationships**: Unlike linear models, decision trees can capture non-linear relationships between variables, which is useful if certain supply chain variables (e.g., distance to hub or flood impact) interact in complex, non-linear ways.
- **Interpretability**: Trees are relatively easy to interpret as they provide visual representations of decision paths. This makes them useful for understanding how specific features drive predictions.

Random forest

- **Improved Accuracy**: Random forest builds an ensemble of decision trees, which reduces overfitting and improves generalization compared to a single tree.
- **Robustness to Variance**: It handles high-dimensional data well and is robust to noise, which can be beneficial if your dataset includes outliers or irrelevant features.
- **Feature Importance**: It provides a measure of feature importance, which helps understand which variables have the most significant impact on warehouse performance or product weight prediction.

Gradient boosting machines

- **Boosting Technique**: GBM builds models sequentially, with each new model trying to correct the errors made by the previous ones. This gradual improvement process often leads to highly accurate predictions in complex datasets like those found in supply chain problems.
- **Handling Non-linearity:** Like other boosting methods, GBM is well-suited for capturing non-linear relationships between variables, such as the impact of distance from the hub and number of distributors on product weight.
- **Customizable Loss Functions:** GBM allows for the use of various loss functions, making it versatile and adaptable to specific needs, like minimizing the error for predicting warehouse product weight.
- **Flexibility:** GBM can be tailored with various parameters, including learning rate, number of trees, and depth, to prevent overfitting while maintaining performance.

XGBoost

- **Efficiency and Performance**: XGB is known for its speed and performance, especially on large datasets. Its boosting technique incrementally builds models, correcting the errors of prior models, leading to better overall accuracy.
- **Handling Non-linear Interactions**: Like random forests, XGB can model complex, non-linear interactions between features. It also incorporates regularization, reducing overfitting.
- **Flexibility**: XGB is versatile and can be tuned extensively to suit the specific nature of the supply chain problem.

Support vector regression

- **High Dimensionality**: SVR is powerful in high-dimensional feature spaces and can model complex relationships by using different kernel functions.
- **Handling Non-linearity**: With kernel tricks (e.g., radial basis function), SVR can handle non-linear relationships between variables like warehouse capacity and transport issues.

- **Regularization**: SVR inherently includes a regularization parameter, controlling the tradeoff between fitting the training data and keeping the model simple, reducing overfitting.

Metrics chosen to validate the model

Accuracy metrics

- R^2 and Adjusted R^2

Adjusted R^2 is better than R^2 since R^2 may have noise.

Error metrics

- MAE, MAPE, MSE, and RMSE

MAPE shows the percentage deviation, is robust to outliers, and is scale-independent.

Model performance

	Model	Set	MAE	MSE	RMSE	MAPE	R^2	Adjusted R^2
0	Linear Regression	Training	1004.414799	1.897114e+06	1377.357690	7.508944	0.985948	0.985926
1	Linear Regression	Test	1008.231535	1.929766e+06	1389.160281	7.697363	0.985609	0.985557
2	Ridge Regression	Training	1004.396897	1.897115e+06	1377.358107	7.507734	0.985948	0.985926
3	Ridge Regression	Test	1008.207852	1.929766e+06	1389.160144	7.696098	0.985609	0.985557
4	Lasso Regression	Training	1004.004205	1.897238e+06	1377.402502	7.503922	0.985947	0.985925
5	Lasso Regression	Test	1007.698287	1.929538e+06	1389.078236	7.691788	0.985611	0.985559
6	Decision Tree	Training	0.000000	0.000000e+00	0.000000	0.000000	1.000000	1.000000
7	Decision Tree	Test	813.273067	1.533688e+06	1238.421721	5.051963	0.988563	0.988522
8	Random Forest	Training	247.390537	1.144128e+05	338.249591	1.527414	0.999153	0.999151
9	Random Forest	Test	666.389379	8.107735e+05	900.429645	4.186479	0.993954	0.993932
10	GBM	Training	670.854269	7.862892e+05	886.729479	4.185065	0.994176	0.994167
11	GBM	Test	675.996468	7.973082e+05	892.921153	4.268707	0.994054	0.994033
12	XGBoost	Training	452.674290	3.546333e+05	595.510936	2.778793	0.997373	0.997369
13	XGBoost	Test	654.653323	7.790656e+05	882.646911	4.177676	0.994190	0.994169
14	SVR	Training	9319.172228	1.280067e+08	11314.002433	76.561152	0.051847	0.050382
15	SVR	Test	9294.730799	1.271062e+08	11274.137795	76.681460	0.052136	0.048711

Model performance after tuning

	Model	Set	MAE	MSE	RMSE	MAPE	R^2	Adjusted R^2
0	Multiple Linear Regression	Training	1008.388852	1.912804e+06	1383.041739	7.553455	0.985782	0.985764
1	Multiple Linear Regression	Test	998.281118	1.874673e+06	1369.186863	7.622025	0.986023	0.985981
2	Tuned Ridge Regression	Training	1008.388061	1.912804e+06	1383.041741	7.553374	0.985782	0.985764
3	Tuned Ridge Regression	Test	998.280924	1.874673e+06	1369.187007	7.621945	0.986023	0.985981
4	Tuned Lasso Regression	Training	1007.968155	1.912838e+06	1383.053923	7.549327	0.985782	0.985764
5	Tuned Lasso Regression	Test	997.752986	1.874349e+06	1369.068787	7.616857	0.986026	0.985983
6	Tuned Decision Tree	Training	576.595835	6.412198e+05	800.761998	3.490106	0.995234	0.995228
7	Tuned Decision Tree	Test	658.570270	8.269075e+05	909.344538	4.087536	0.993835	0.993816
8	Tuned Random Forest	Training	218.433544	9.633935e+04	310.385802	1.348227	0.999284	0.999283
9	Tuned Random Forest	Test	566.911702	6.476952e+05	804.795138	3.540163	0.995171	0.995156
10	Tuned Gradient Boosting Machine	Training	358.603404	2.326026e+05	482.288893	2.208899	0.998271	0.998269
11	Tuned Gradient Boosting Machine	Test	582.104715	6.544867e+05	809.003540	3.690032	0.995120	0.995106
12	Tuned XGBoost	Training	341.286640	2.094035e+05	457.606218	2.093345	0.998444	0.998442
13	Tuned XGBoost	Test	586.103035	6.577793e+05	811.035944	3.716356	0.995096	0.995081
14	Tuned Support Vector Regression	Training	2093.533848	9.539827e+06	3088.661100	14.885744	0.929092	0.929000
15	Tuned Support Vector Regression	Test	2129.966172	9.848088e+06	3138.166351	15.156588	0.926577	0.926354

Key insights

- **Tuning Effects**: Tuning improved most models, especially in reducing errors and enhancing R^2. The most significant improvements were seen in tree-based models (Decision Tree, Random Forest, GBM, XGBoost).
- **Model Performance**: XGBoost and Random Forest consistently outperformed other models, even after tuning, making them the most reliable for this dataset. XGBoost, in particular, showed superior performance post-tuning.
- **Overfitting**: Initial overfitting in the Decision Tree was mitigated through tuning, resulting in better generalization.
- **Significant features:**

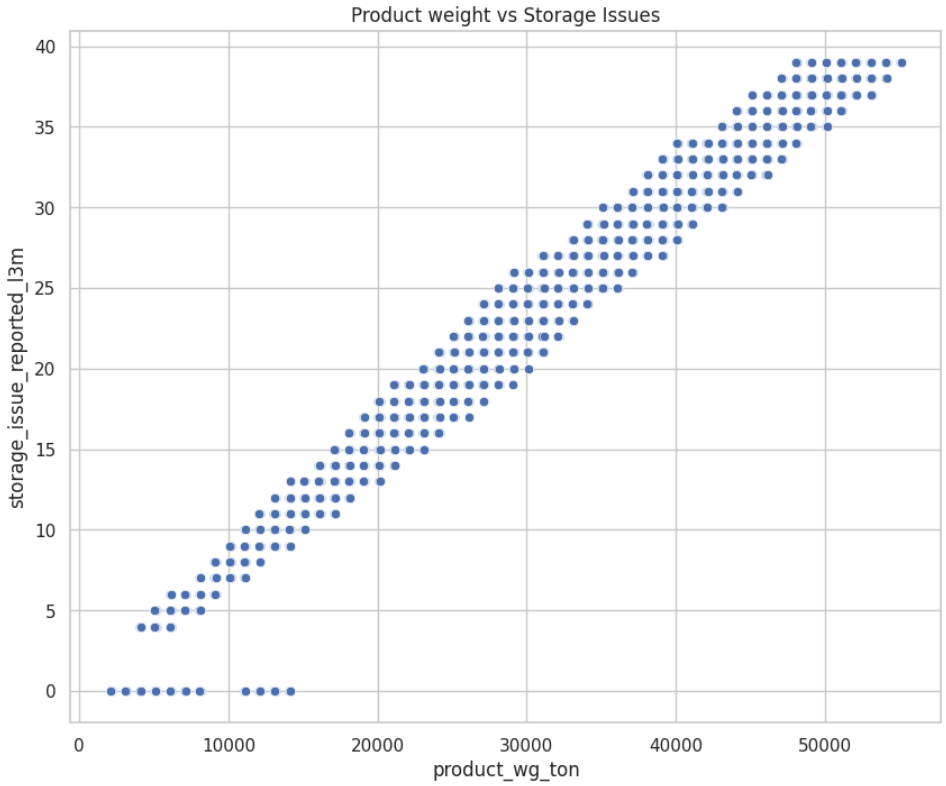
Positive contributors (increase product weight):

	Feature	Coefficient
7	storage_issue_reported_l3m	11674.487113
26	temp_reg_mach_1	196.912869
17	approved_wh_govt_certificate_A+	118.610351
20	approved_wh_govt_certificate_C	111.061976
0	num_refill_req_l3m	50.935521
16	zone_West	27.436611
4	distributor_num	26.495038
13	WH_capacity_size_Small	19.753562
5	dist_from_hub	16.581413
22	wh_owner_type_Rented	12.077306
14	zone_North	7.627735
9	govt_check_l3m	6.545316
24	flood_proof_1	4.710661
3	retail_shop_num	3.884640

Negative contributors (decrease product weight):

	Feature	Coefficient
6	workers_num	-1.680533
15	zone_South	-2.114927
12	WH_capacity_size_Mid	-2.266460
23	flood_impacted_1	-3.992453
25	electric_supply_1	-8.629265
2	Competitor_in_mkt	-24.092691
21	Location_type_Urban	-24.601096
11	cluster	-72.848777
10	warehouse_age	-218.187120
8	wh_breakdown_l3m	-320.371332
1	transport_issue_l1y	-369.927068
18	approved_wh_govt_certificate_B	-813.362277
19	approved_wh_govt_certificate_B+	-829.210697

Most significant feature: Storage issue reported in the last 3 months



We can also observe a strong positive correlation between the target variable product weight and the variable storage issue reported in last three months.

Significant Positive Contributors:

- Storage Issue Reported Last 3 Months
- Temperature Regulation Machines
- Government Certification A+ and C
- Zone (North, South, West): North, South, and West
- Number of Refill Requests Last 3 Months

Significant Negative Contributors:

- Government Certification B and B+
- Warehouse Age
- Transport Issues
- Warehouse Breakdown
- Location Type Urban

Multicollinearity:

	Feature	VIF
0	num_refill_req_l3m	1.102372
1	transport_issue_l1y	2.352618
2	Competitor_in_mkt	1.271753
3	retail_shop_num	1.040640
4	distributor_num	1.001993
5	dist_from_hub	1.002710
6	workers_num	1.158613
7	storage_issue_reported_l3m	2.219821
8	wh_breakdown_l3m	1.274175
9	govt_check_l3m	1.153864
10	warehouse_age	1.764836
11	cluster	2.885403
12	WH_capacity_size_Mid	1.275982
13	WH_capacity_size_Small	1.284958
14	zone_North	16.110005
15	zone_South	13.251778
16	zone_West	13.982935
17	approved_wh_govt_certificate_A+	1.875895
18	approved_wh_govt_certificate_B	1.649244
19	approved_wh_govt_certificate_B+	1.654878
20	approved_wh_govt_certificate_C	1.844446
21	Location_type_Urban	1.008040
22	wh_owner_type_Rented	1.075455
23	flood_impacted_1	1.054221
24	flood_proof_1	1.022825
25	electric_supply_1	1.193920
26	temp_reg_mach_1	1.378388

- **Zone-related features (zone_North, zone_South, zone_West)** have very high VIF values (above 10), indicating **strong multicollinearity**. These features are likely redundant in the model, and you might need to drop or combine them to avoid multicollinearity issues.

- Most of the other features have **VIF values below 5**, which indicates that multicollinearity is not a significant problem for them. They can be safely retained in the model.

Overall, tree-based models, especially ensemble methods like Random Forest and XGBoost, provided the best performance both before and after tuning.

Interpretation of the Most Optimum Model: XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning method that builds multiple trees sequentially, with each tree trying to correct the errors made by the previous one. It effectively handles both bias and variance, making it highly accurate and robust.

Performance Metrics:

- **Training MAE:** 601.47
- **Test MAE:** 636.57
- **Training MSE:** 644,110.8
- **Test MSE:** 735,773.0
- **Training R^2:** 0.9952
- **Test R^2:** 0.9945

These metrics indicate that the model fits the data very well, with a minimal difference between training and test performance, suggesting strong generalization ability.

Business implications:

- The high R^2 value (0.9945) means that the XGBoost model explains approximately 99.45% of the variance in the target variable, making it highly reliable for predictions.
- The low MAE (636.57 on the test set) and MSE demonstrate that the model's predictions are close to the actual values, minimizing potential errors in decision-making.
- Accurate predictions allow for better strategic planning and resource allocation. For instance, if the model is used to predict sales, production needs, or customer demand, the business can adjust its operations to optimize costs, inventory levels, and service delivery.
- XGBoost is efficient and scalable, meaning it can handle large datasets and complex relationships within the data. As the business grows and accumulates more data, the

model can continue to provide reliable predictions without a significant drop in performance.

- The model's predictions can help optimize resource allocation by forecasting demand more accurately, thus reducing waste and improving efficiency. This can lead to cost savings and better utilization of resources across various departments.

Recommendations

- **Enhance Temperature Regulation Capabilities**
  - Advanced temperature control systems can be put in place.
- **Leverage Strategic Location Clusters:**
  - North, South, West show increased product weight.
  - Optimize supply routes to ensure faster and more efficient deliveries to and from these locations.
- **Address Storage and Capacity Issues:**
  - Optimize space utilization, implement better inventory management practices such as just-in-time inventory.
  - Deploy advanced storage solutions like vertical racking systems to maximize available space
- **Invest in Warehouses with A+ and C Certifications:**
  - Upgrade existing B and B+ certified warehouses to higher standards by improving operational processes, quality checks, and infrastructure to achieve A+ or C certification levels.
- **Address Transport Issues and Breakdown Prevention:**
  - Use predictive analytics to anticipate breakdowns and address them before they disrupt operations.
- **Optimize Warehouse Age and Modernize Infrastructure:**
  - Implement newer technologies like automated storage and retrieval systems.
  - Gradually phase out older warehouses and replace them with newer, more efficient facilities