CAPSTONE PROJECT

ON

SUPPLY CHAIN OPTIMIZATION

PREPARED

BY

KARTHIKEYAN M.P.

Business Problem:

1.Introduction :

a)Defining problem statement

A FMCG company has entered into the instant noodles business two years back. Their higher management has notices that there is a miss match in the demand and supply, leading to inventory cost loss to the company.

b) Need of the study/project

The management wants to minimize the inventory cost loss by optimizing the supply quantity in each and every warehouse in entire country. This project is intended to build a model to predict the optimum weight of product to be shipped to each warehouse.

c) Understanding business/social opportunity

The FMCG company by predicting the optimized weight for supplying to the warehouses can reduce the inventory cost loss. Streamlining the supplychain will inturn increase the customer base and the profitability. By analysing the demand patterns in different pockets of country the company can plan curated marketing strategies so as to engage with existing customers and acquire new customers.

2. EDA and Business Implication

- Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?

Sample Dateset

Table 1.1 Sample Dataset

	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt
0	WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	2
1	WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	4
2	WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	4
3	WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	2
4	WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	2
5	WH_100005	EID_50005	Rural	Small	West	Zone 1	8	0	2
6	WH_100006	EID_50006	Rural	Large	West	Zone 6	8	0	4
7	WH_100007	EID_50007	Rural	Large	North	Zone 5	1	0	4
8	WH_100008	EID_50008	Rural	Small	South	Zone 6	8	1	4
9	WH_100009	EID_50009	Rural	Small	South	Zone 6	4	3	3

Shape

The dataset has 25000 rows and 24 columns.

Understanding of attributes

Variable info

Variables	Non null values	Datatype
Ware_house_ID	25000	object
WH_Manager_ID	25000	object
Location_type	25000	object
WH_capacity_size	25000	object
zone	25000	object
WH_regional_zone	25000	object
num_refill_req_l3m	25000	int 64
transport_issue_l1y	25000	int 64
Competitor_in_mkt	25000	int 64
retail_shop_num	25000	int 64
wh_owner_type	25000	object
distributor_num	25000	int 64
flood_impacted	25000	int 64
flood_proof	25000	int 64
electric_supply	25000	int 64
dist_from_hub	25000	int 64
workers_num	24010	float 64
wh_est_year	13119	float 64
storage_issue_reported_l3m	25000	int 64
temp_reg_mach	25000	int 64
approved_wh_govt_certificate	24092	object
wh_breakdown_l3m	25000	int 64
govt_check_l3m	25000	int 64
product_wg_ton	25000	int 64

The given dataset has
14 Integer datatype variables
2 Float datatype variables and
8 Object datatype variables

Missing Values

The dataset has missing values in following variables

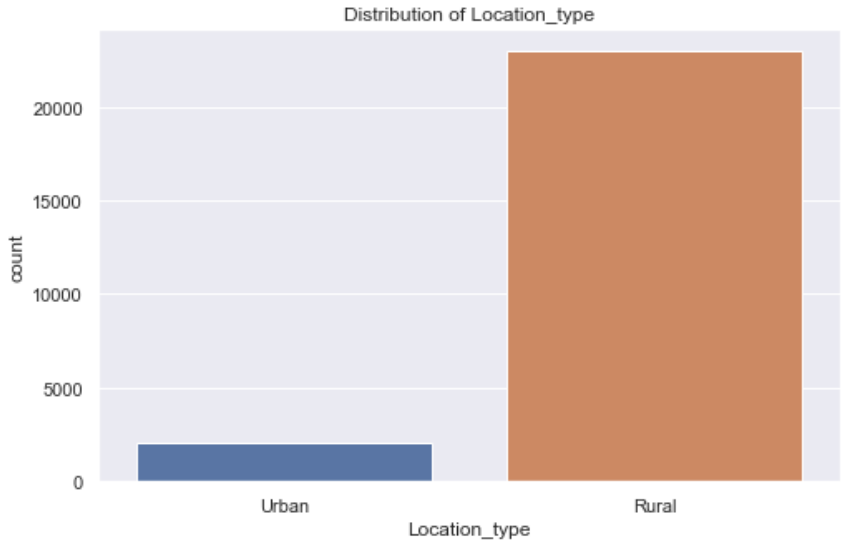
Variables	No. of Missing Values
workers_num	990
wh_est_year	11881
approved_wh_govt_certificate	908

Description of Dataset

Table -1.2. Description of Dataset

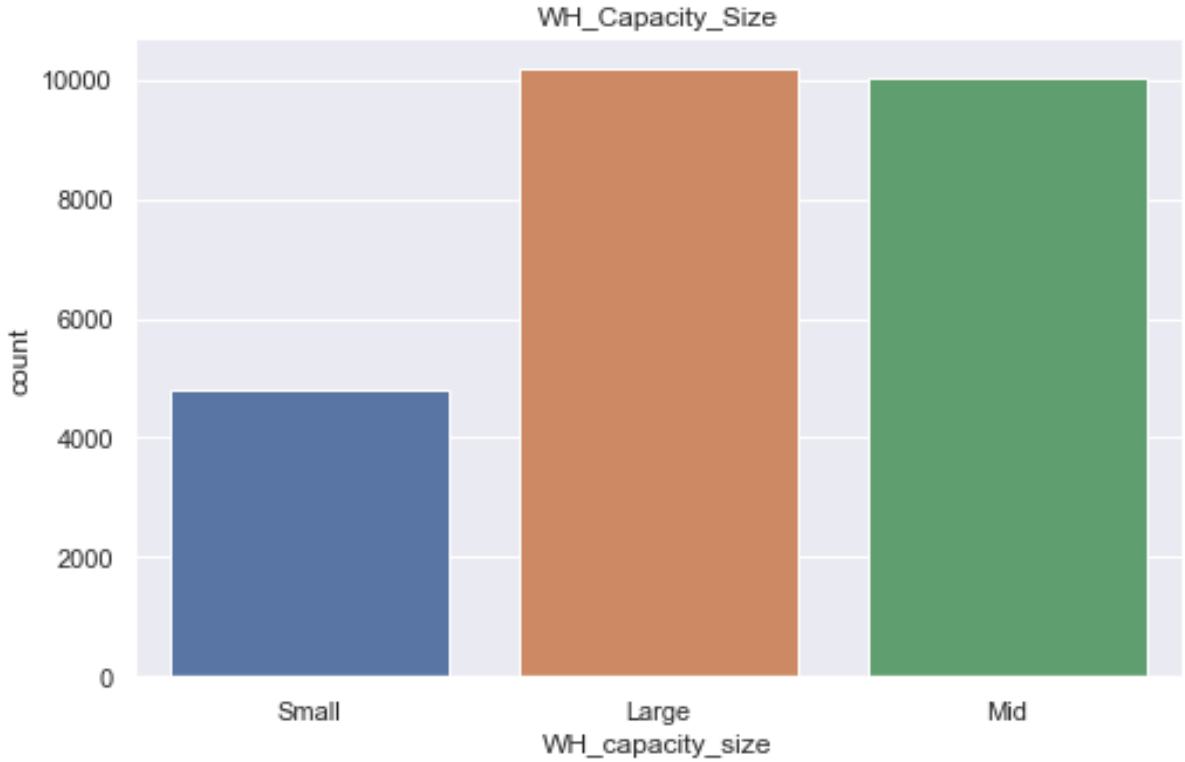
	count	mean	std	min	25%	50%	75%	max
num_refill_req_l3m	25000.0	4.089040	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	0.773680	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	3.104200	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	4985.711560	1052.825252	1821.0	4313.0	4859.0	5500.0	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	25000.0	28.907000	7.717275	10.0	24.0	28.0	33.0	98.0
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
wh_age_years	13119.0	13.616815	7.528230	0.0	7.0	14.0	20.0	27.0

a) Univariate analysis



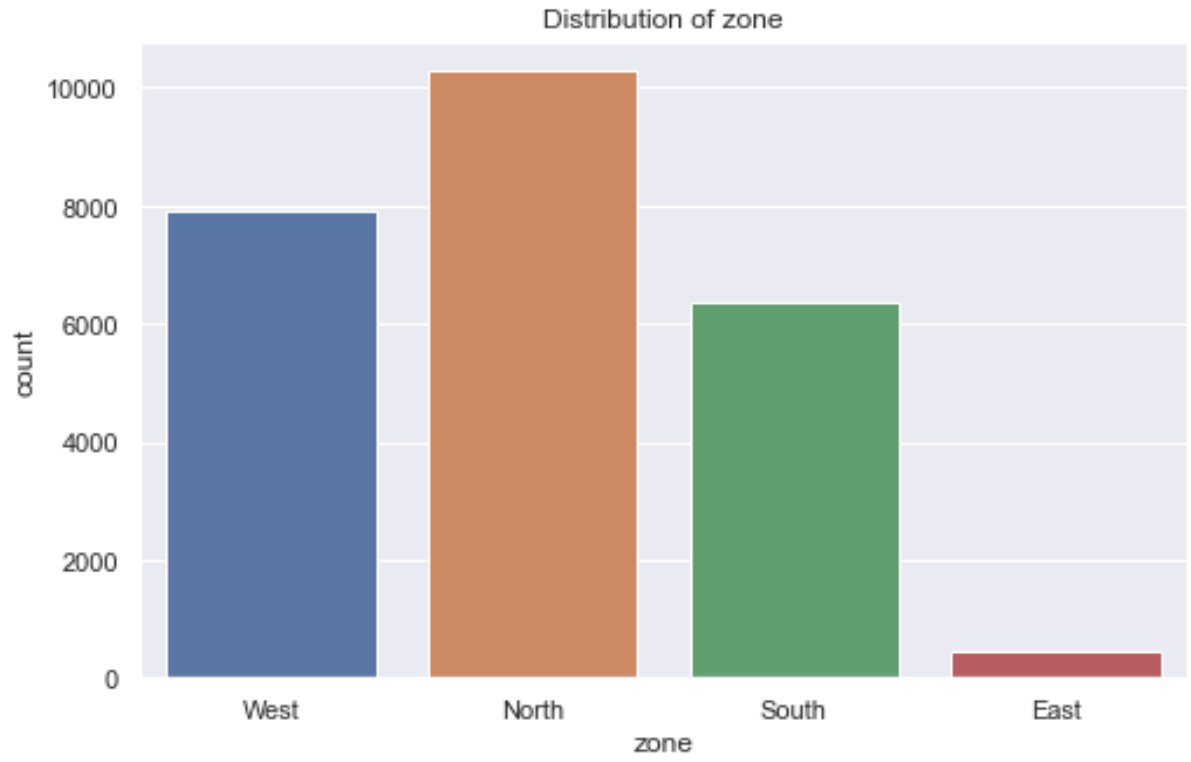
Insight:

Rural locations have higher concentration of warehouses than Urban locations.
Total Rural warehouses = **22957**
Total Urban Warehouses =**2043**



Insight:

Most available warehouses are of Large size followed by medium & small sizes.
Total Large warehouses = 10169
Total Mid Warehouses = 10020
Total Small Warehouses = 4811



Insight:

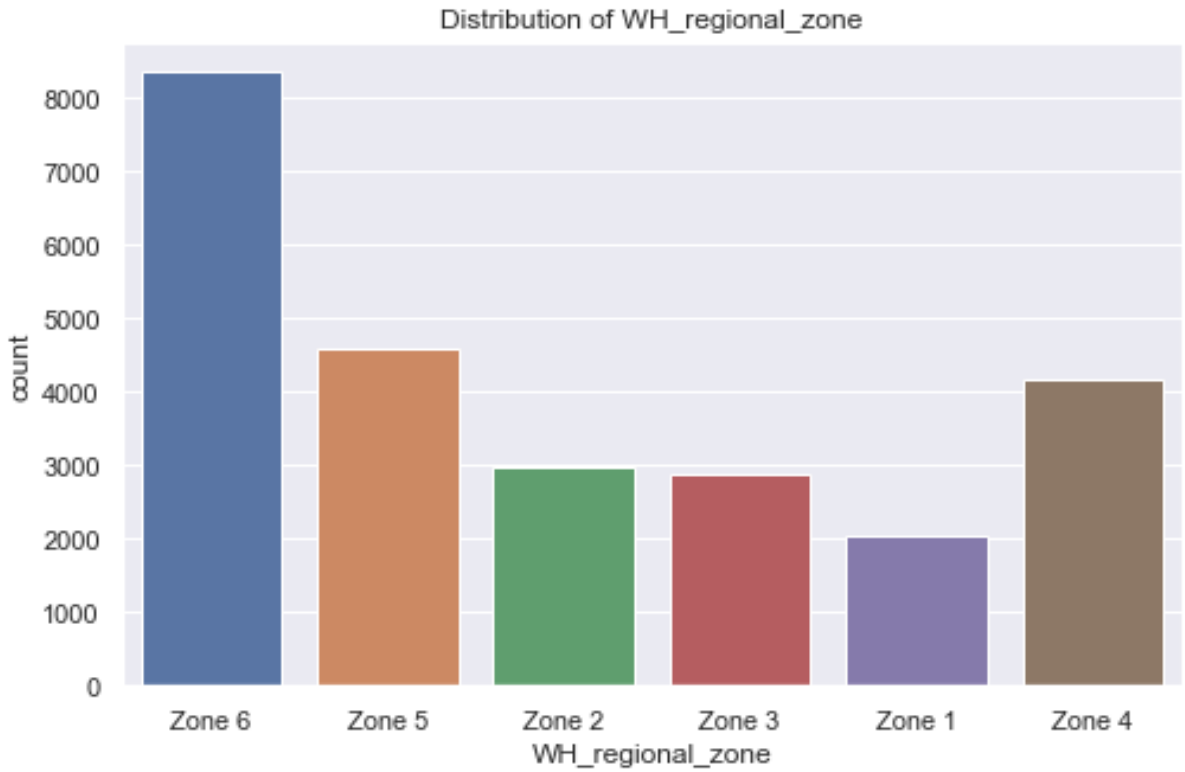
North zone has highest number of warehouses & East has least warehouses.

Total warehouses in North zone= 10278 – (41.11%)

Total Warehouses in South zone = 6362 - (25.44%)

Total Warehouses in West zone = 7931 - (31.72%)

Total Warehouses in East zone = 429 - (1.71%)



Insight:

zone 6 has highest number of warehouses & zone 1 has least warehouses.

Total warehouses in zone 6 = 8339 (33.4%)

Total warehouses in zone 5 = 4587 (18.4%)

Total warehouses in zone 4 = 4176 (18.3%)

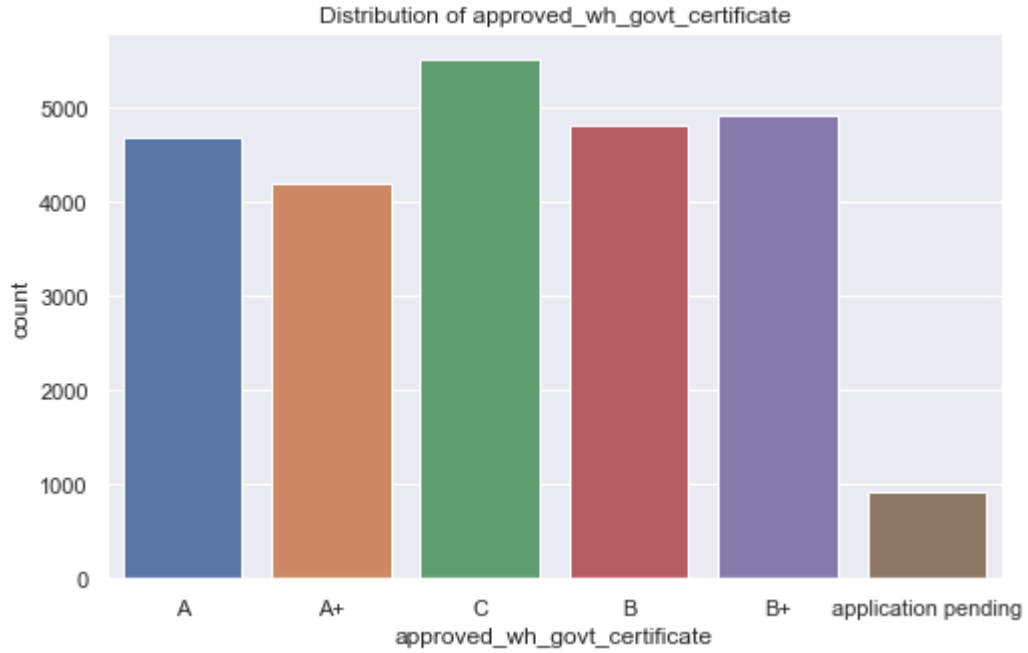
Total warehouses in zone 3 = 2881 (11.5%)

Total warehouses in zone 2 = 2963 (16.7%)

Total warehouses in zone 1 =2054 (8.21%)

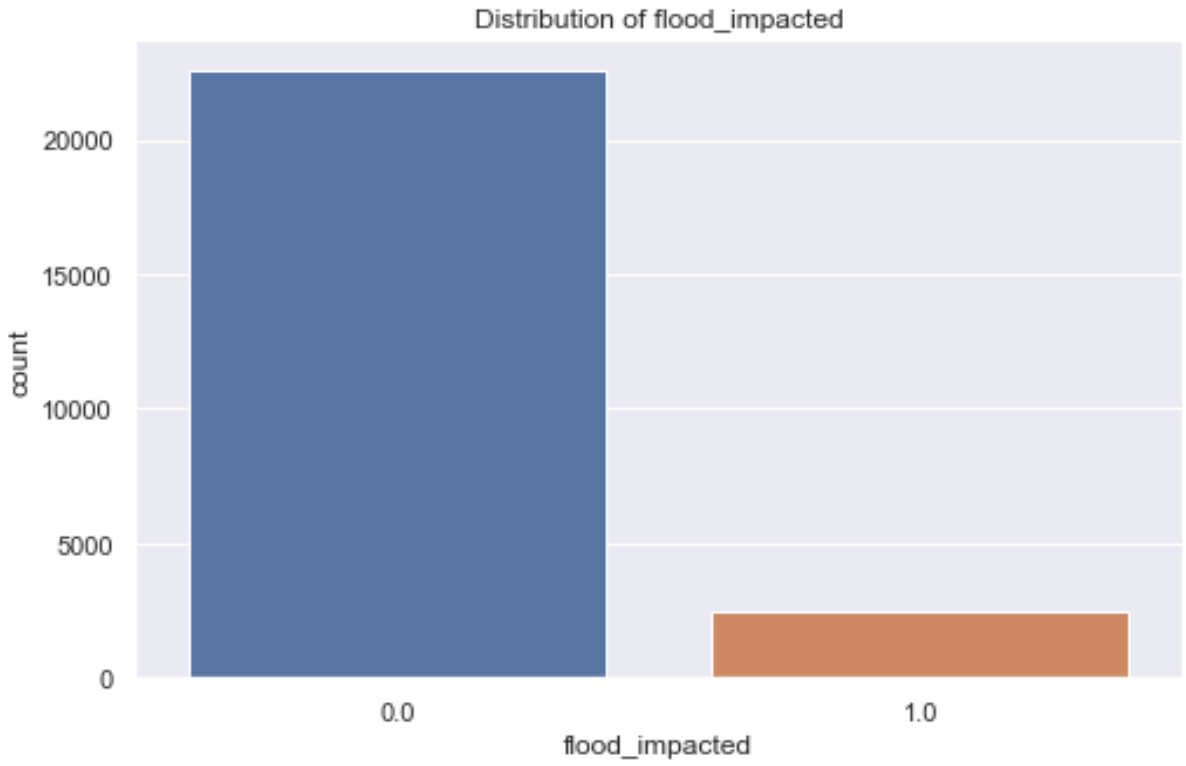
Insight:

Most of the warehouses are Company Owned
Total warehouse owned by Company= 13578 (54.3%)
Total warehouse Rented by Company= 11422 (45.7%)

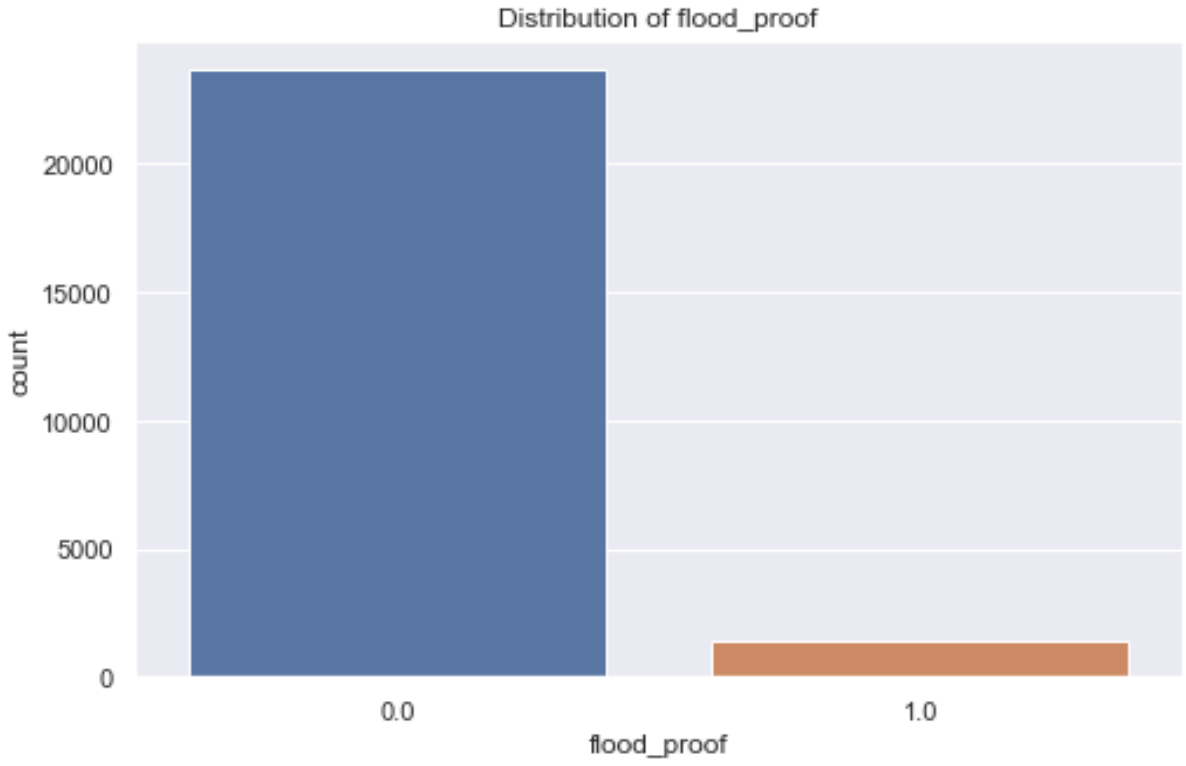


Insight:

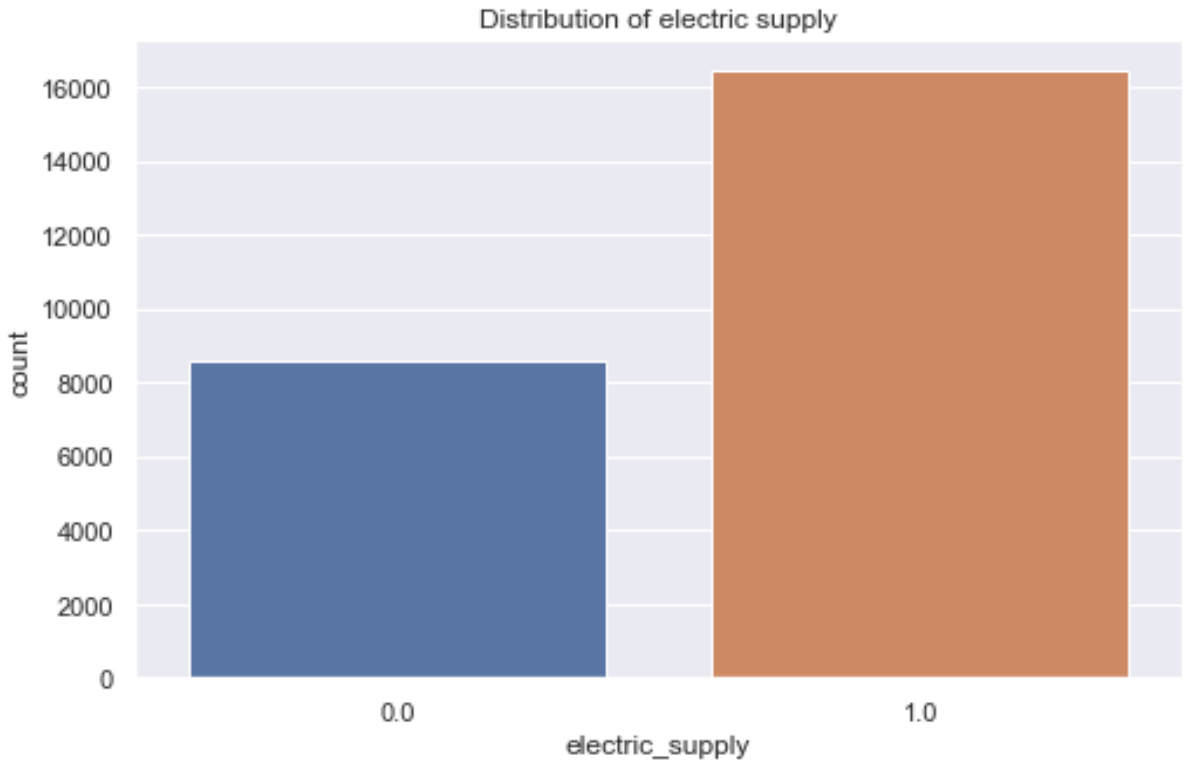
Most of the warehouses have ‘C’ certificate and comparatively least warehouses have ‘A+’



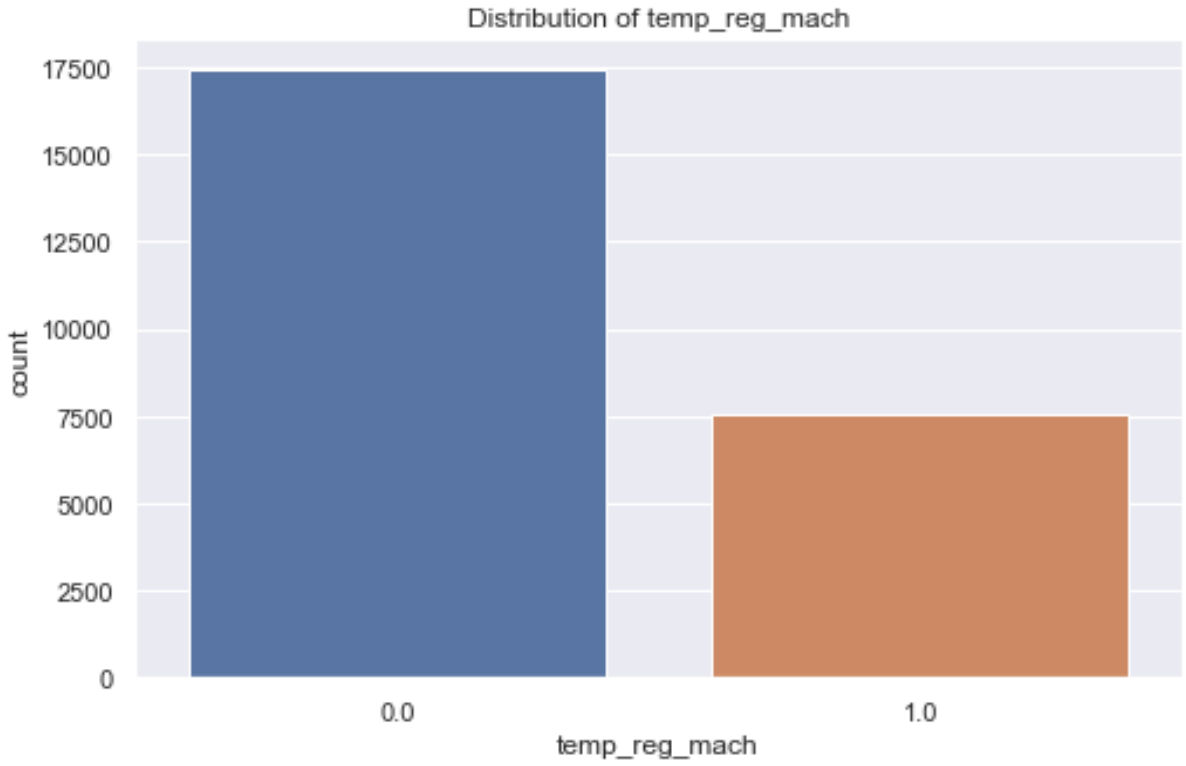
Insight:
9.82% of total warehouse i.e 2454 warehouses are impacted by flood.



Insight:
1366 warehouses are flood proof.



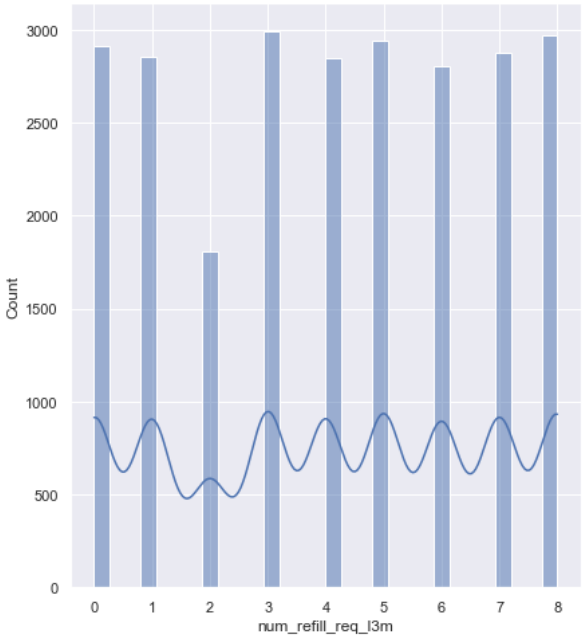
Insight:
16422 warehouses are equipped with backup for power supply while **8578** warehouses lack such power backups.

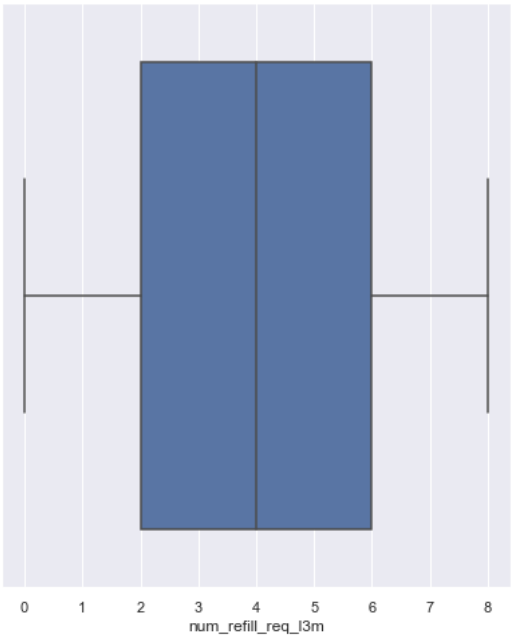


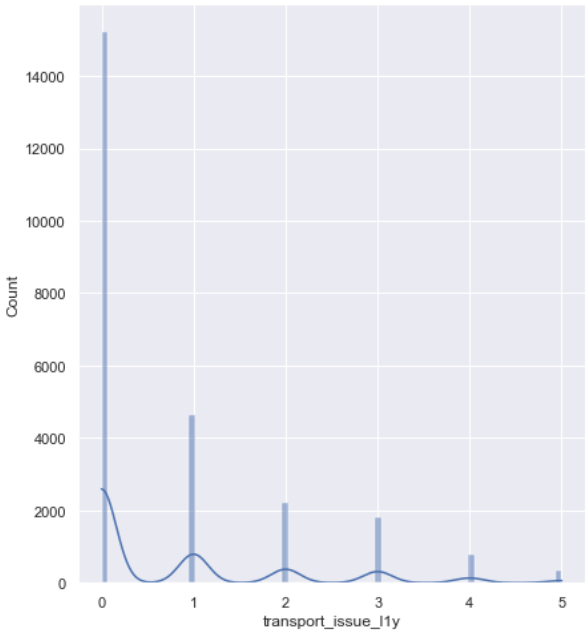
Insight:

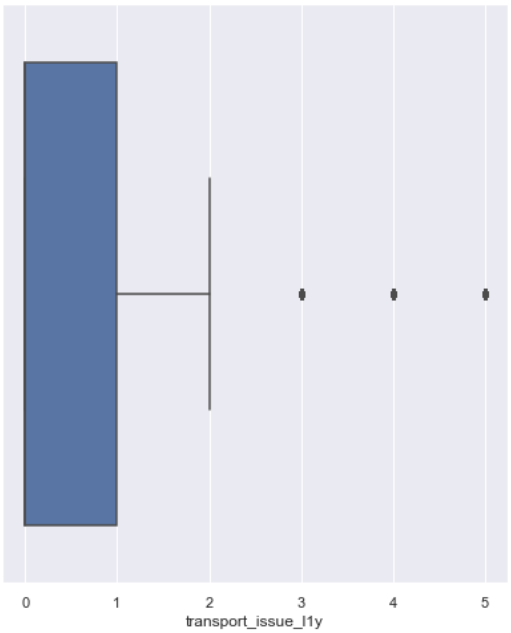
17418 warehouses are installed with temperature regulating machine indicator while 7582 warehouses doesn't have one.

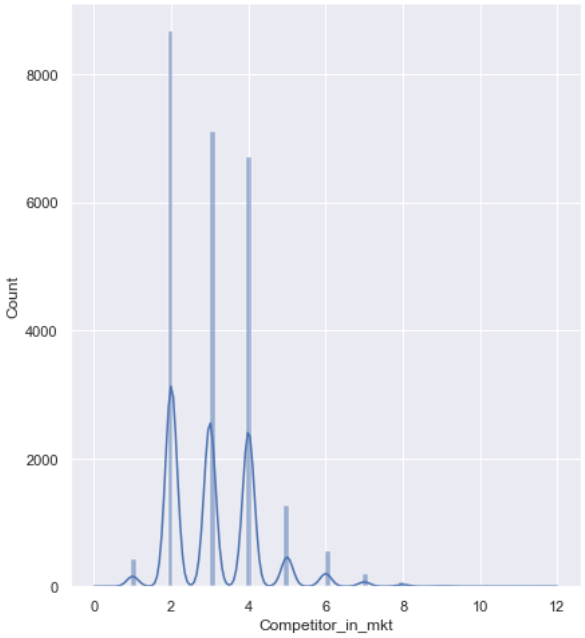
Distribution of Continuous Variables

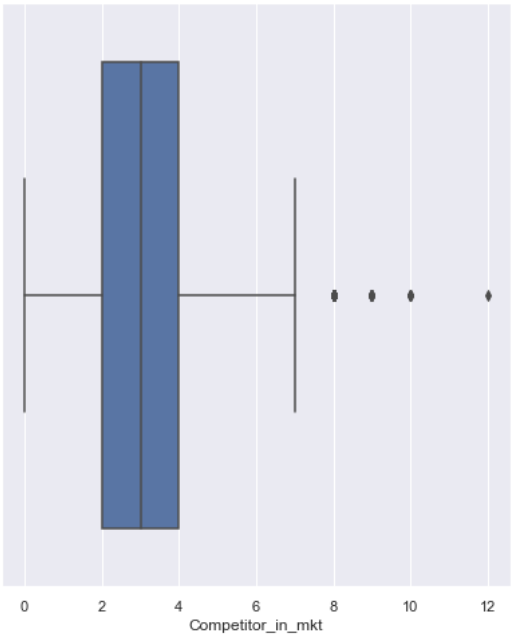


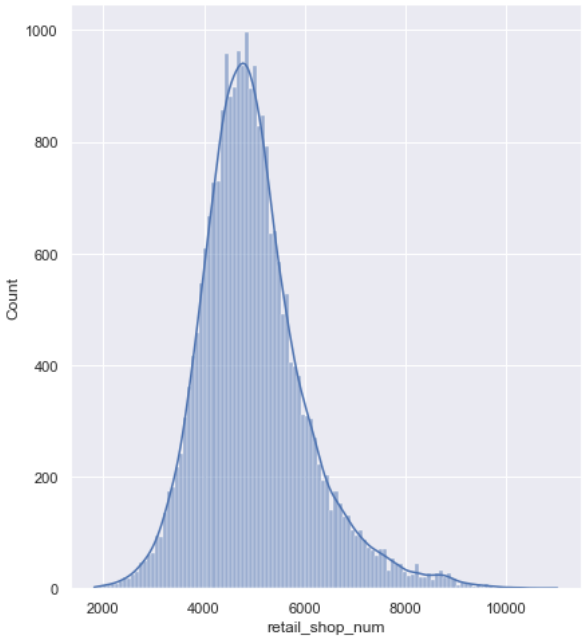


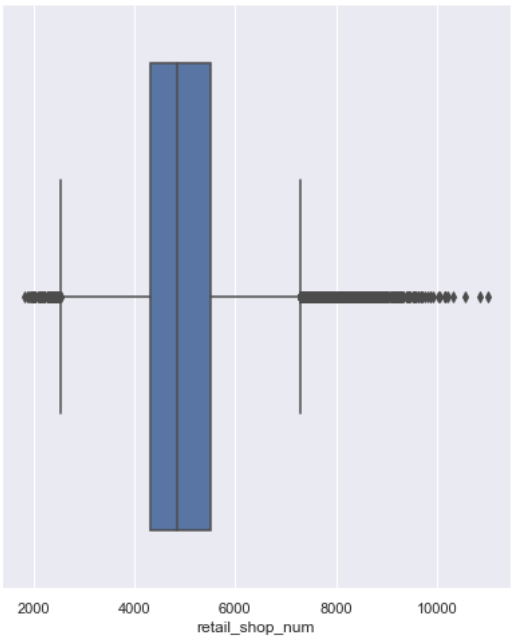


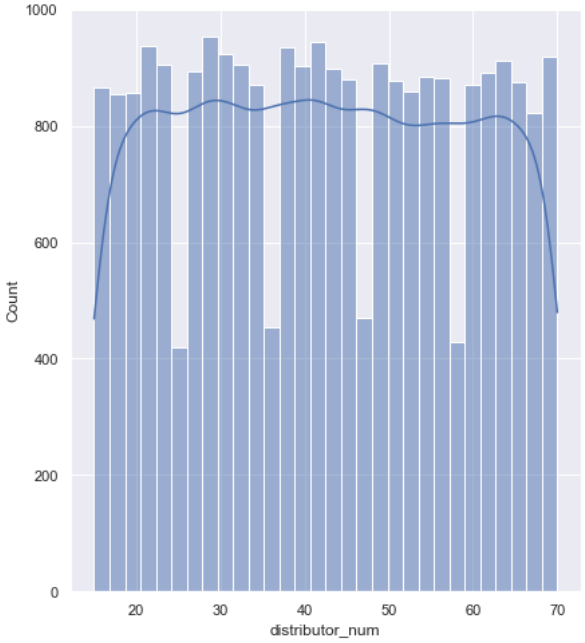


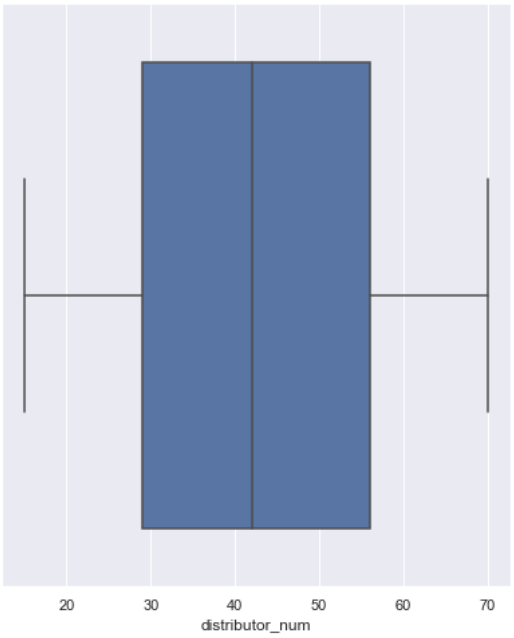

Count
Competitor_in_mkt

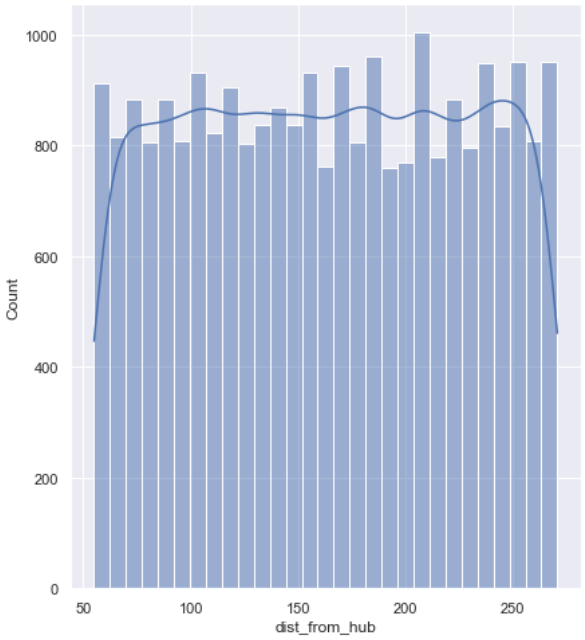

Competitor_in_mkt

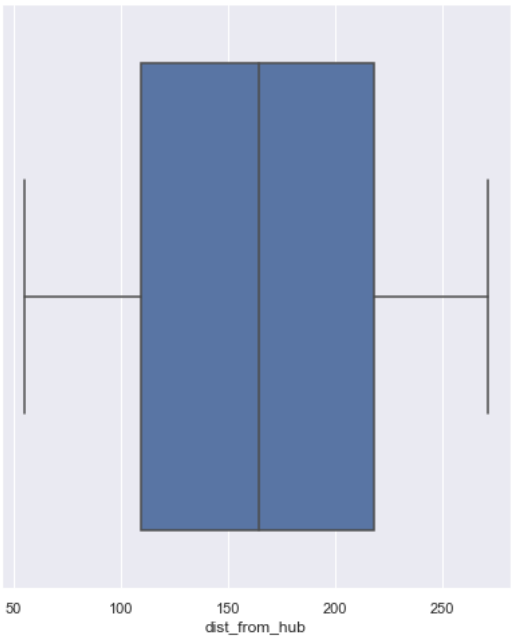

Count
retail_shop_num

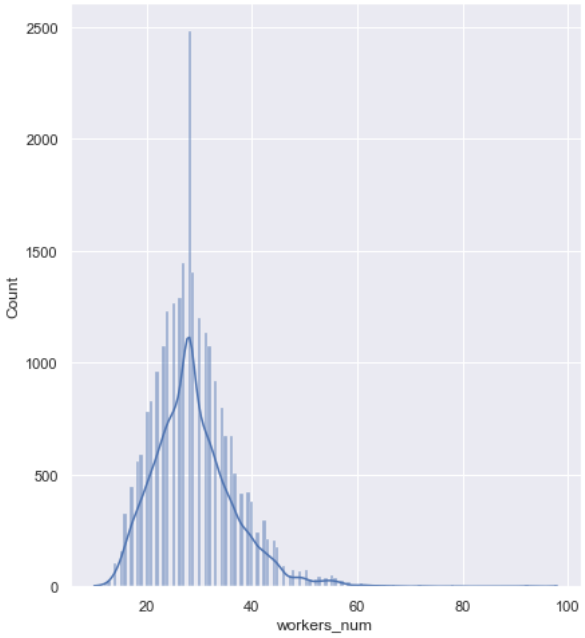

retail_shop_num

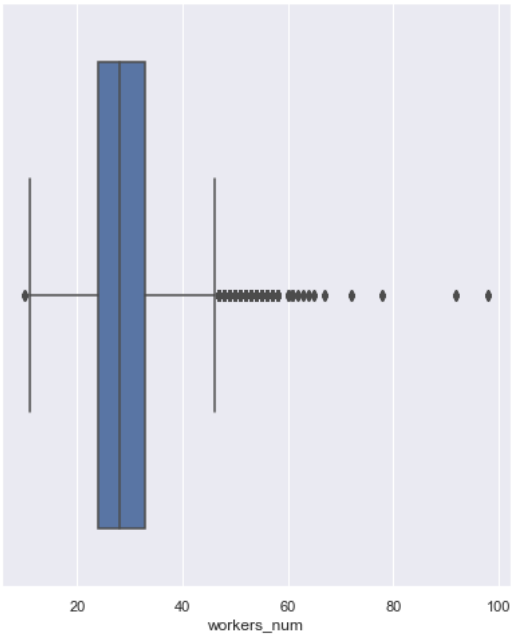

1000
800
600
400
200
0
Count
20
30
40
50
60
70
distributor_num

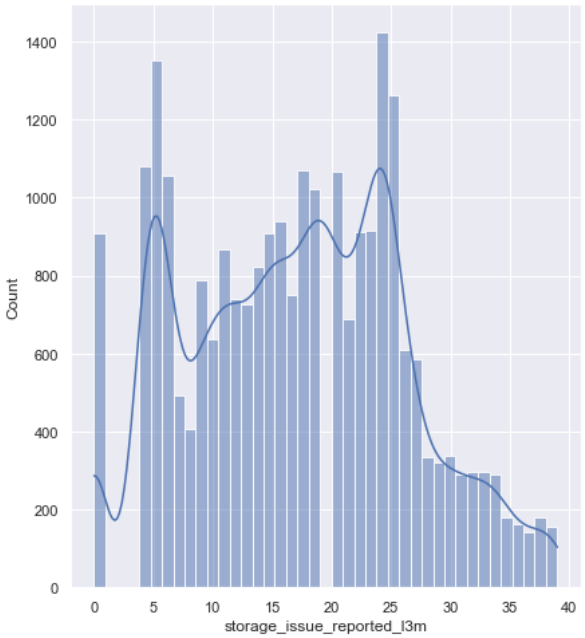

20
30
40
50
60
70
distributor_num

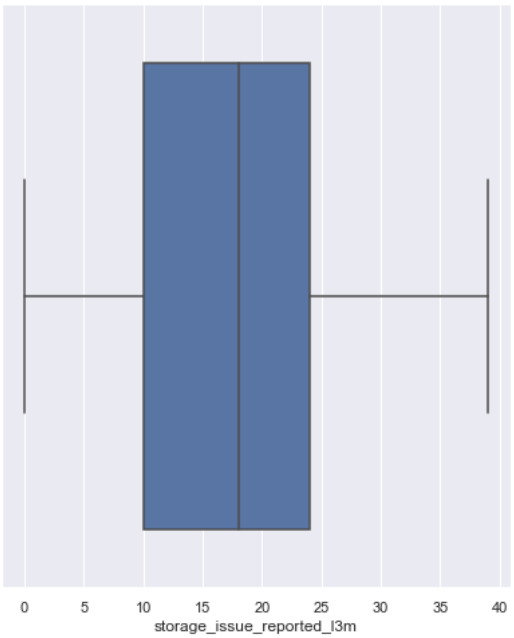

1000
800
600
400
200
0
Count
50
100
150
200
250
dist_from_hub

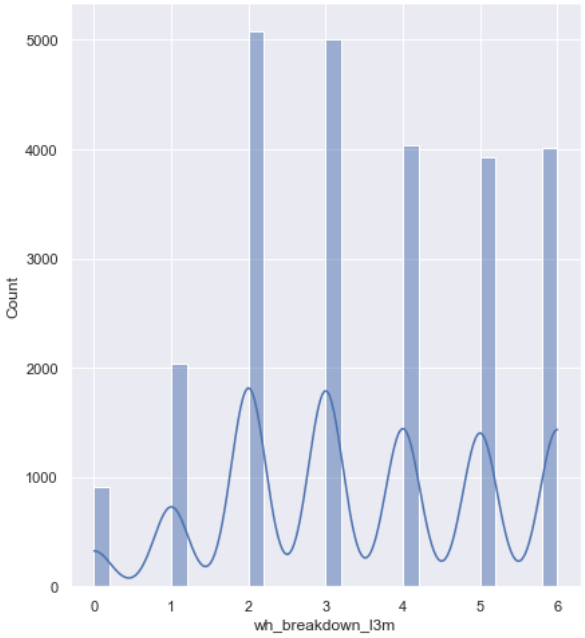

50
100
150
200
250
dist_from_hub

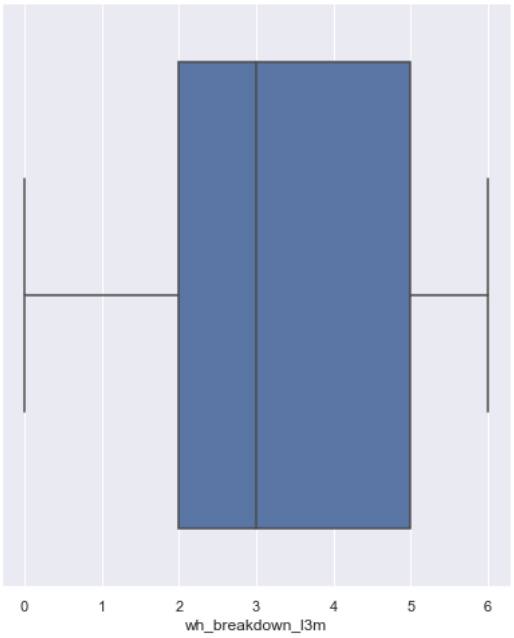

Count
workers_num

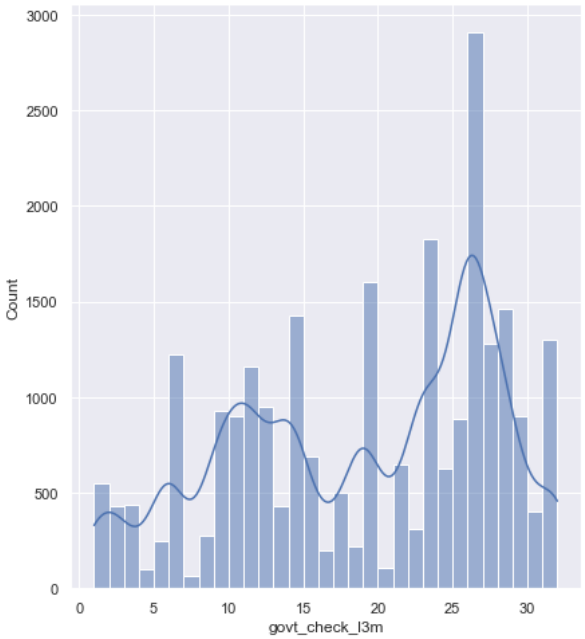

workers_num

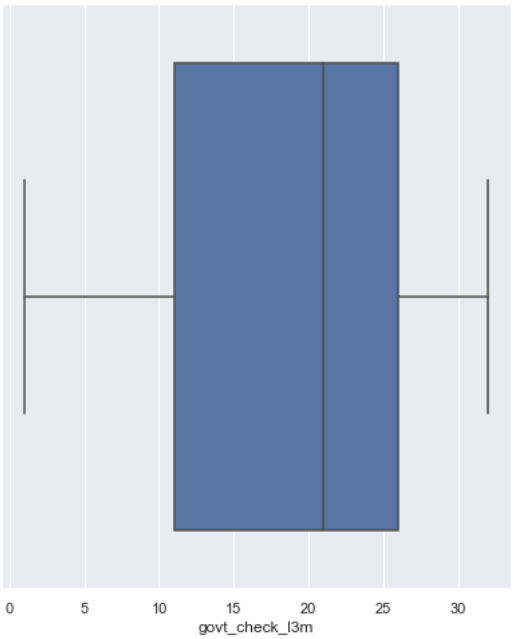

Count
storage_issue_reported_l3m

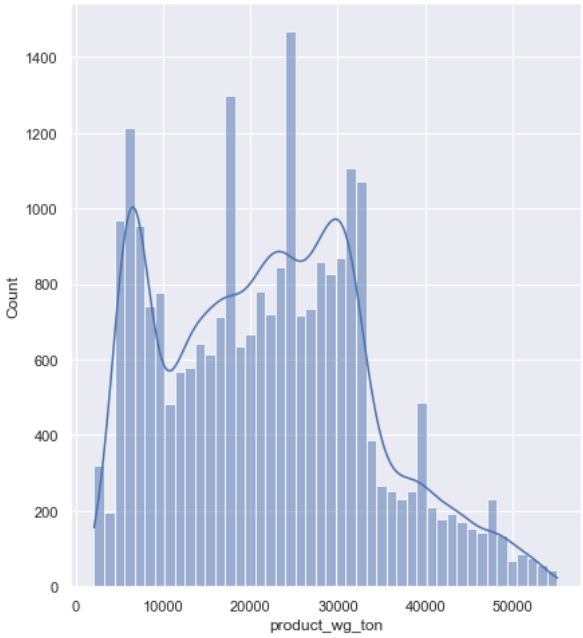

storage_issue_reported_l3m

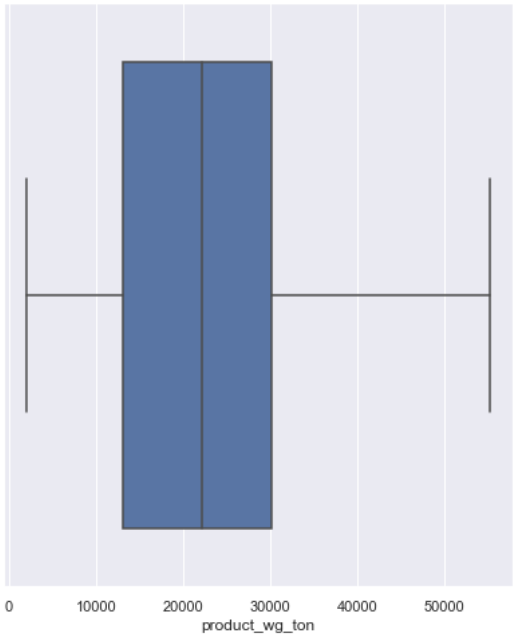

Count
wh_breakdown_l3m

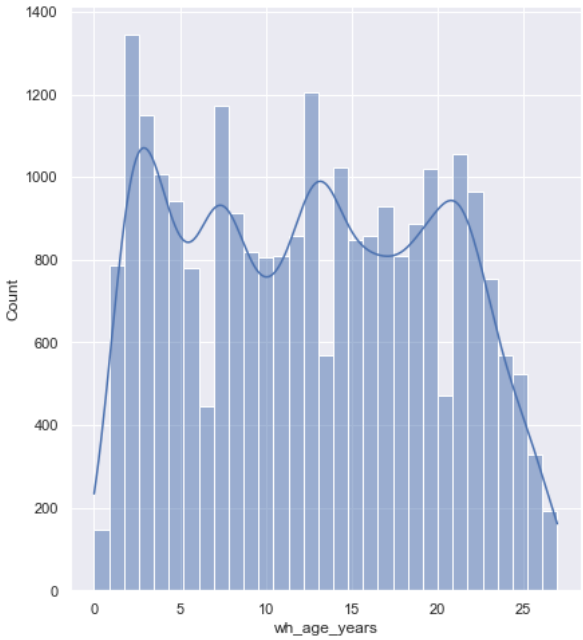

wh_breakdown_l3m

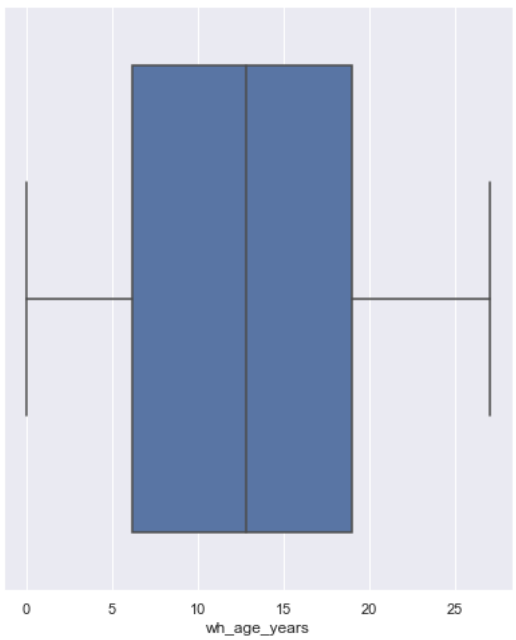

Count
govt_check_l3m


govt_check_l3m



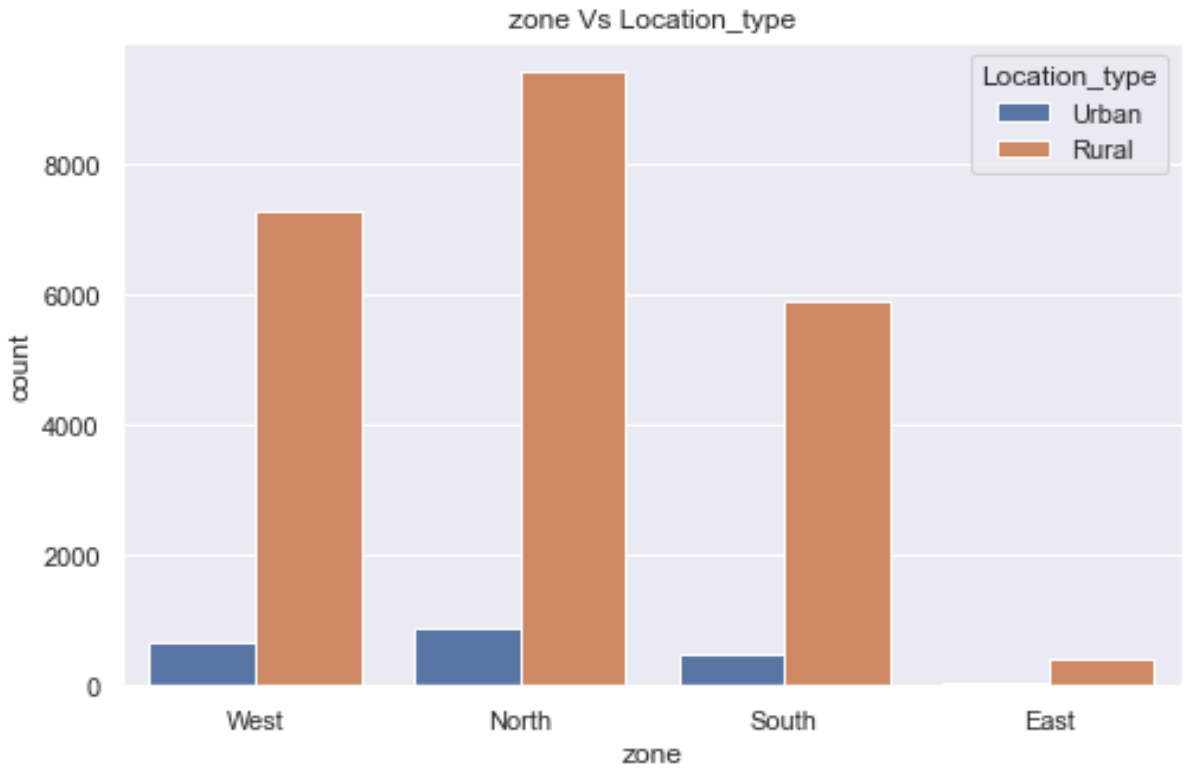






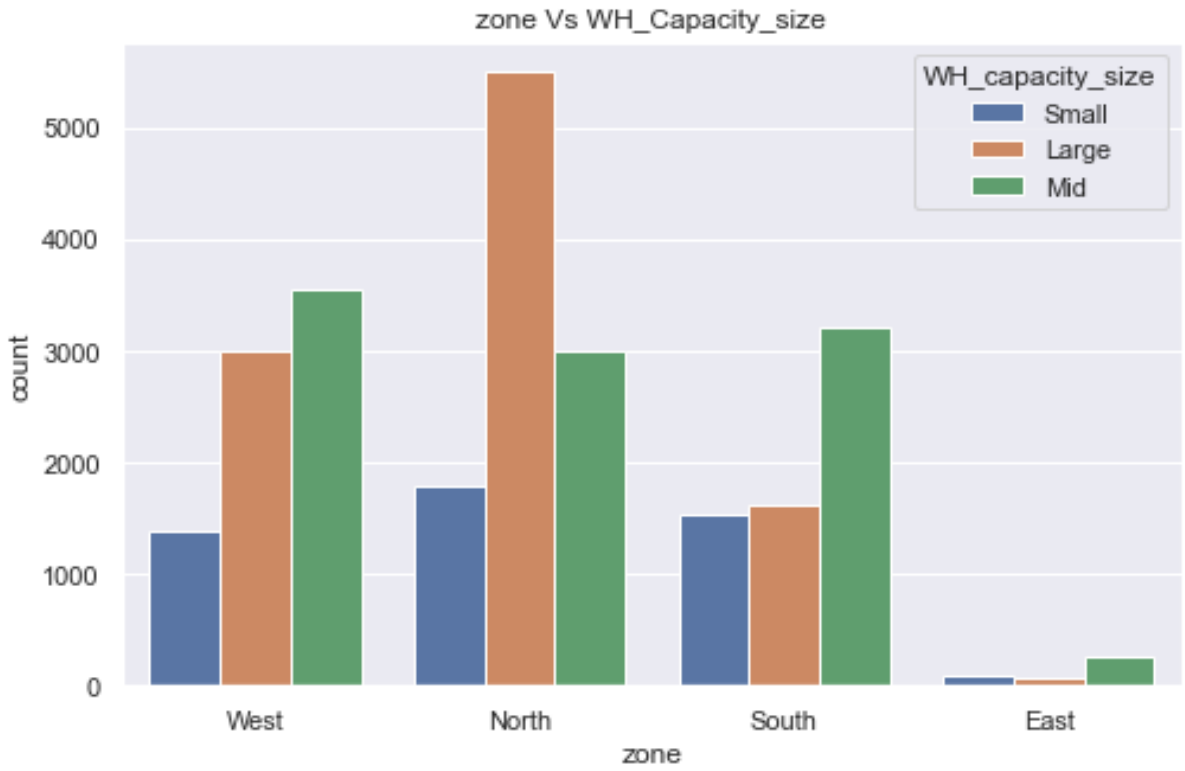
Outlier Detection

The Presence of outliers are detected from boxplots of Variables such as
1.transport_issue_l1y
2.Competitor_in_market
3.workers_num
4.retail_shop_num



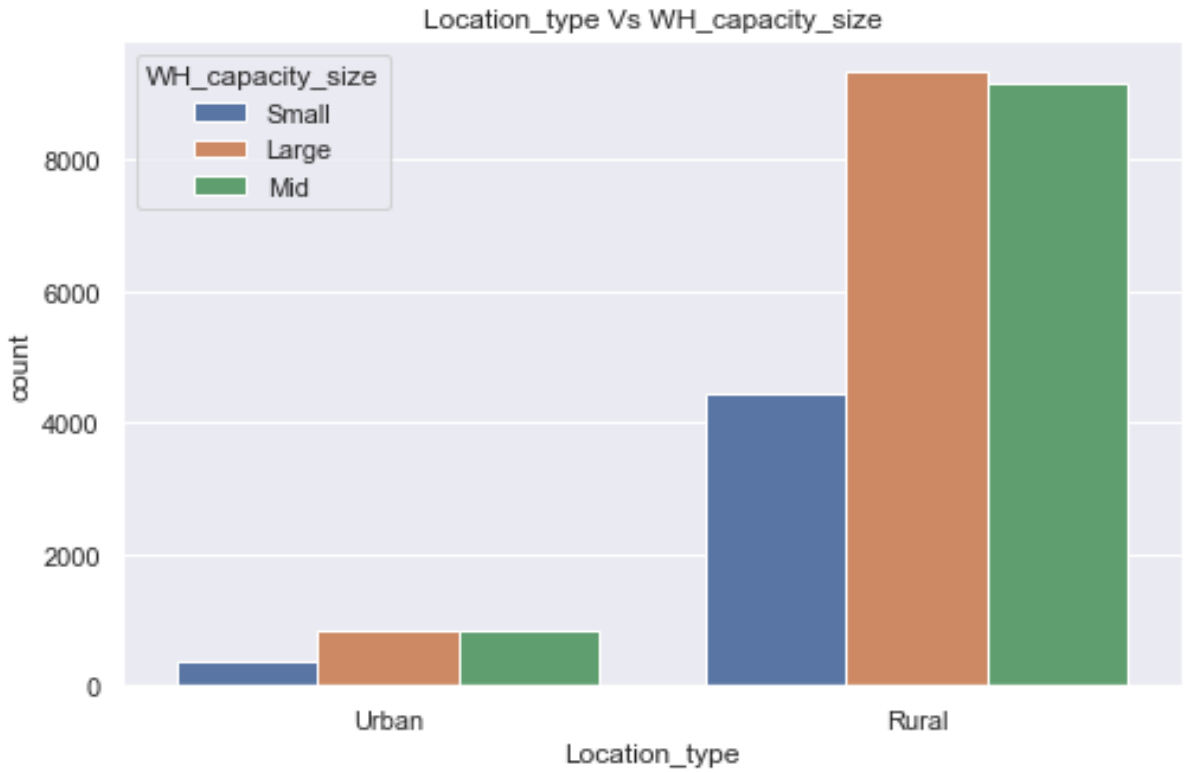
Location_type zone	Rural	Urban	All
East	0.02	0.00	0.02
North	0.38	0.04	0.41
South	0.24	0.02	0.25
West	0.29	0.03	0.32
All	0.92	0.08	1.00

Insight:
Higher concentration of warehouses are in North zone , accounting to 41% of total warehouses



WH_capacity_size	Large	Mid	Small	All
zone				
East	0.00	0.01	0.00	0.02
North	0.22	0.12	0.07	0.41
South	0.06	0.13	0.06	0.25
West	0.12	0.14	0.06	0.32
All	0.41	0.40	0.19	1.00

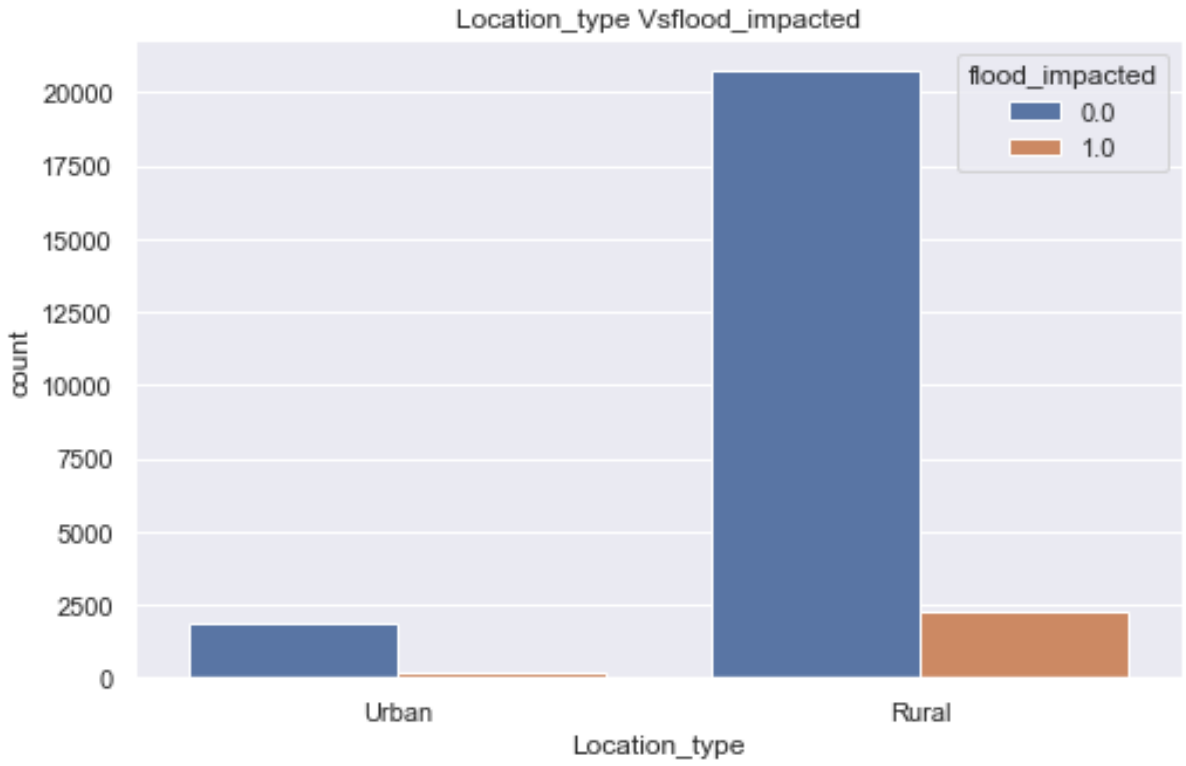
Insight:
Of total warehouses Most warehouses are Large sized accounting 41% followed by Mid sized warehouses 40% and small sized for 19%.



WH_capacity_size	Large	Mid	Small	All
Location_type				
Rural	0.37	0.37	0.18	0.92
Urban	0.03	0.03	0.01	0.08
All	0.41	0.40	0.19	1.00

Insight :

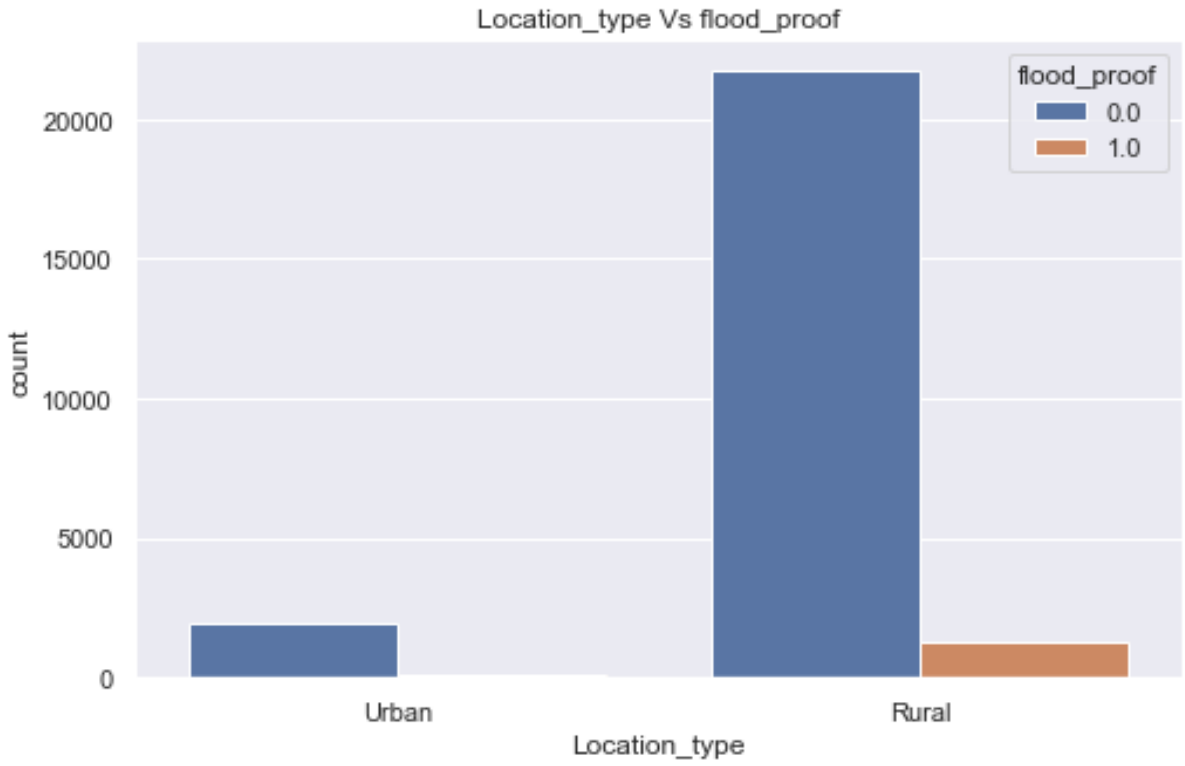
Urban location has Mid warehouses slightly higher than no. of Large size warehouses



flood_impacted	0.0	1.0	All
Location_type			
Rural	0.83	0.09	0.92
Urban	0.07	0.01	0.08
All	0.90	0.10	1.00

Insight :

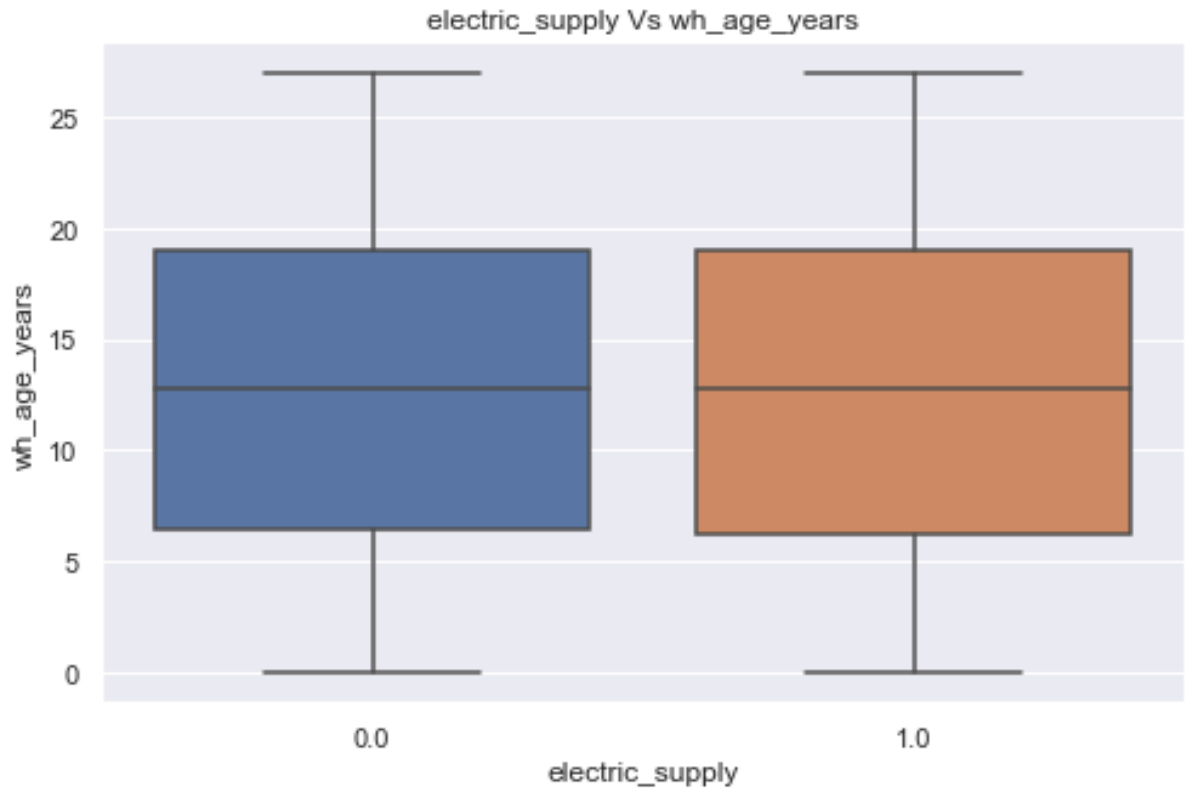
10% of warehouses are impacted by the flood in the past,most of the impacted warehouses are in Rural regions.



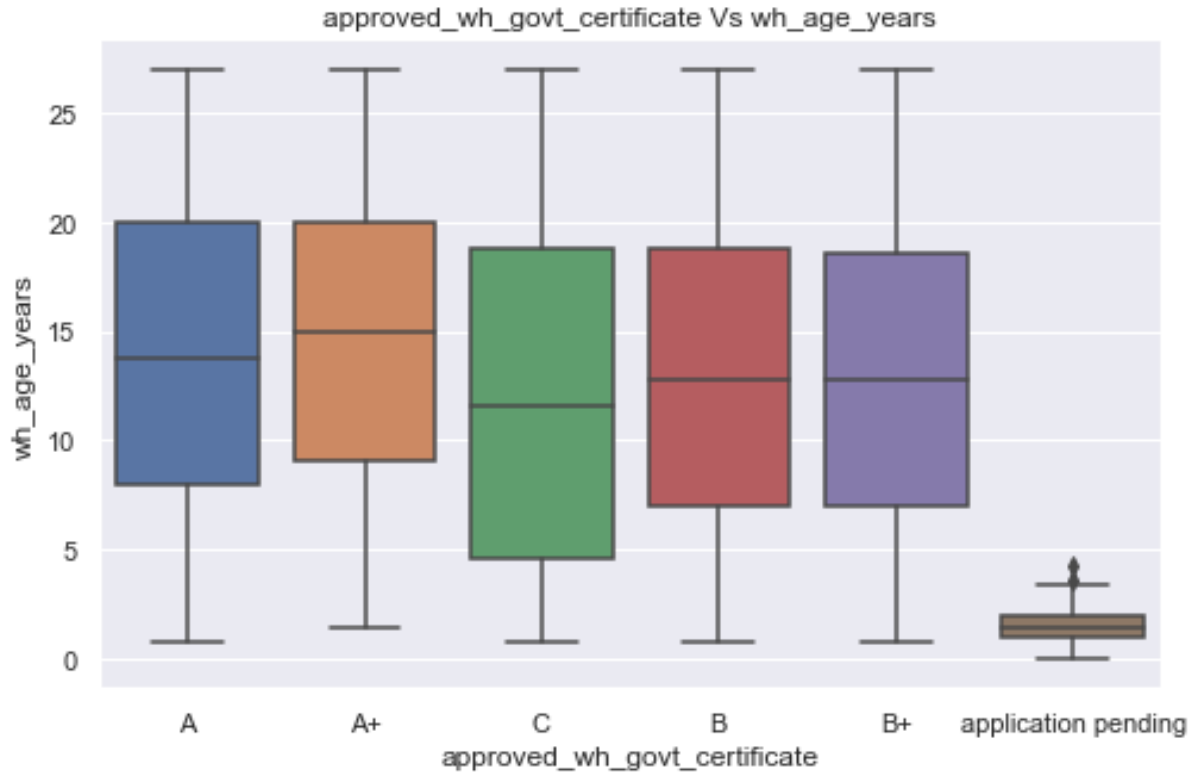
flood_proof	0.0	1.0	All
Location_type			
Rural	0.87	0.05	0.92
Urban	0.08	0.00	0.08
All	0.95	0.05	1.00

Insight :

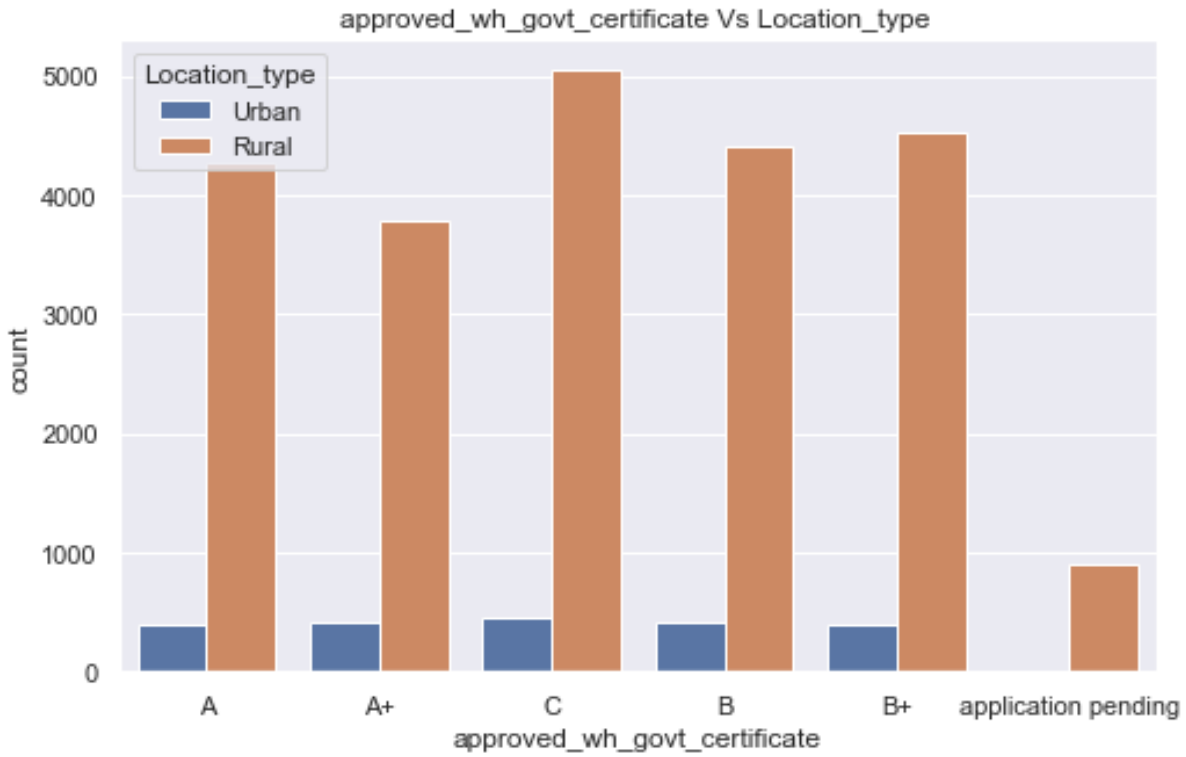
5% of warehouses are flood proof and mostly are in Rural regions



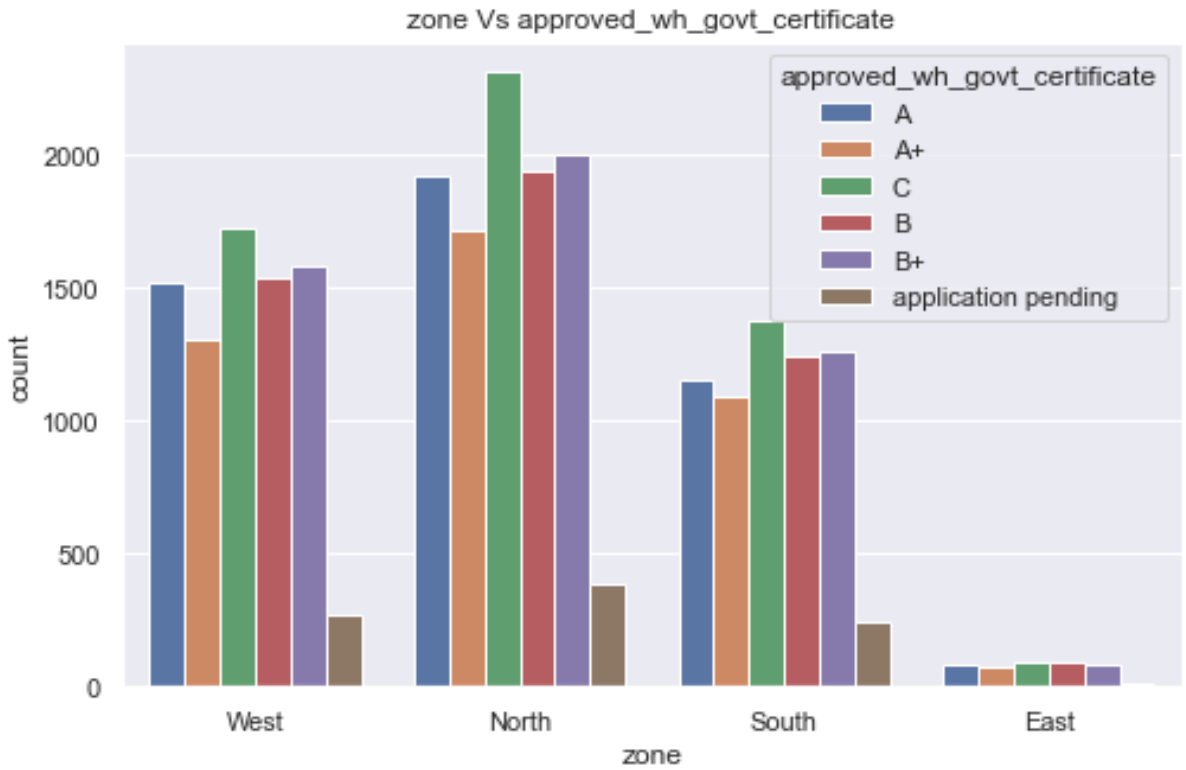
There is lack of power backups found even in some older warehouses



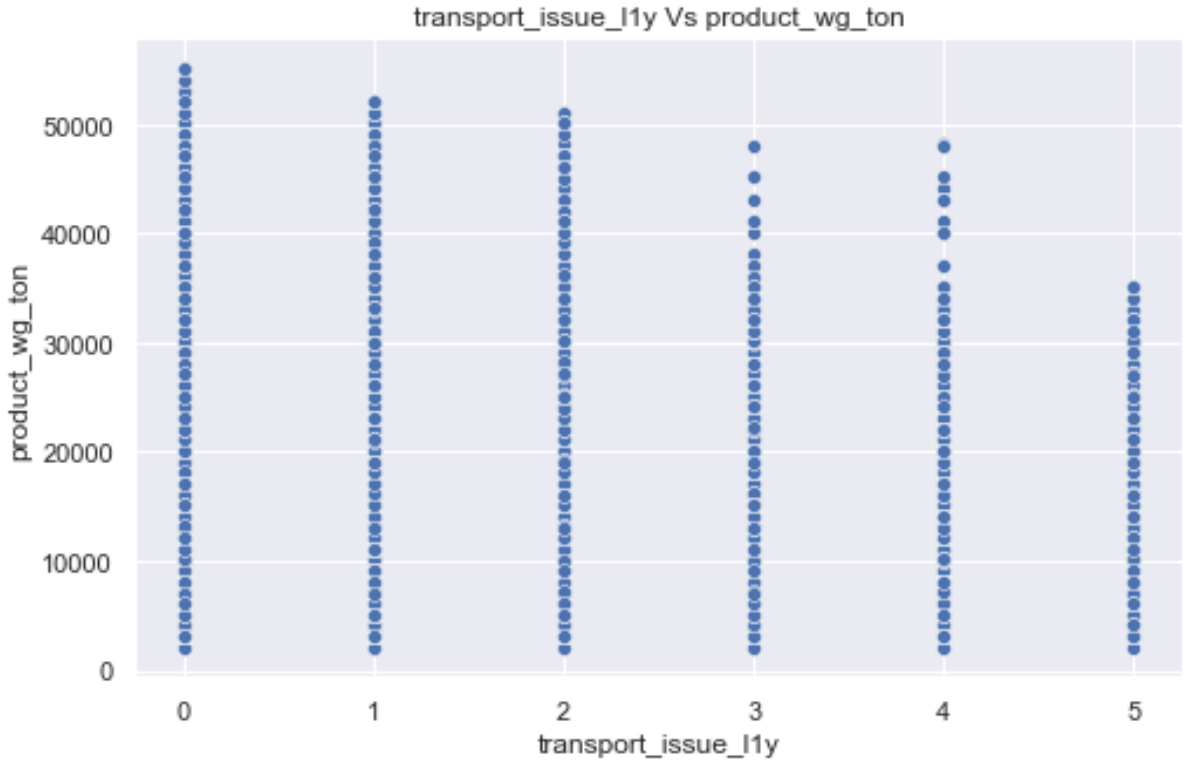
The newly established warehouses with age lesser than 5 years are yet to receive certifications from the government.



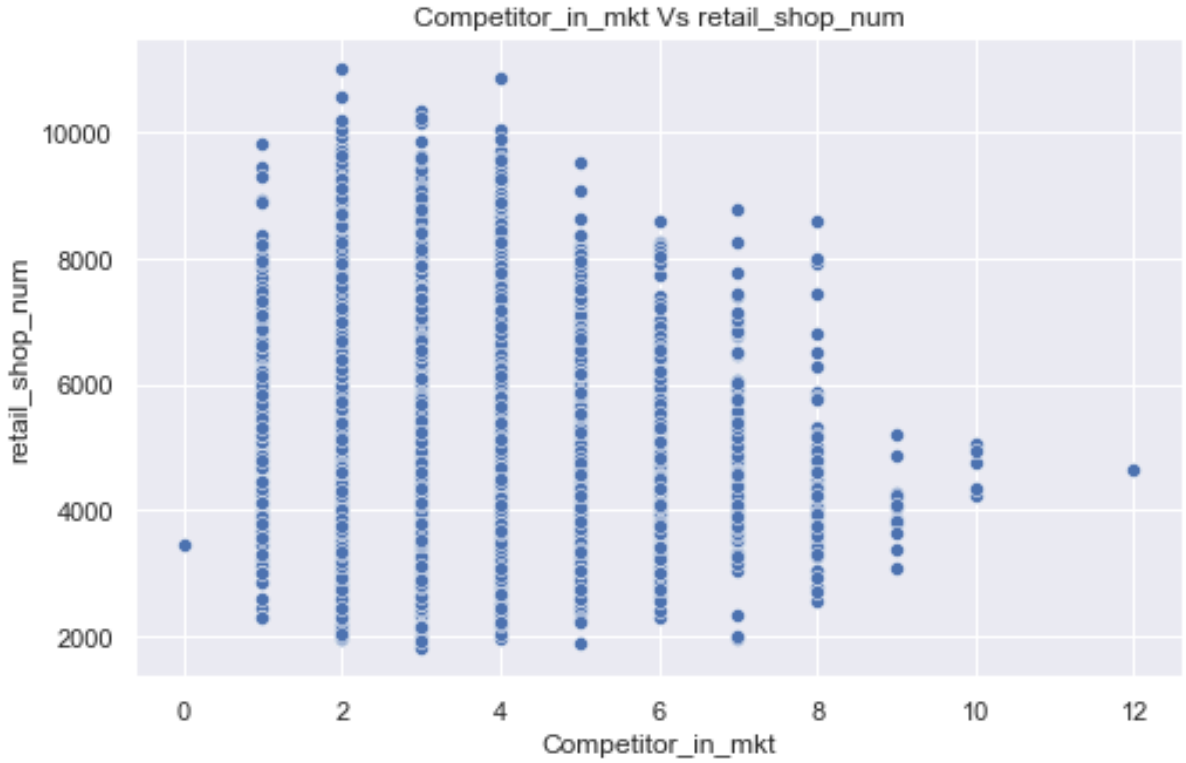
All the warehouses in the Urban areas are certified and some new warehouses in rural areas are yet to be certified.



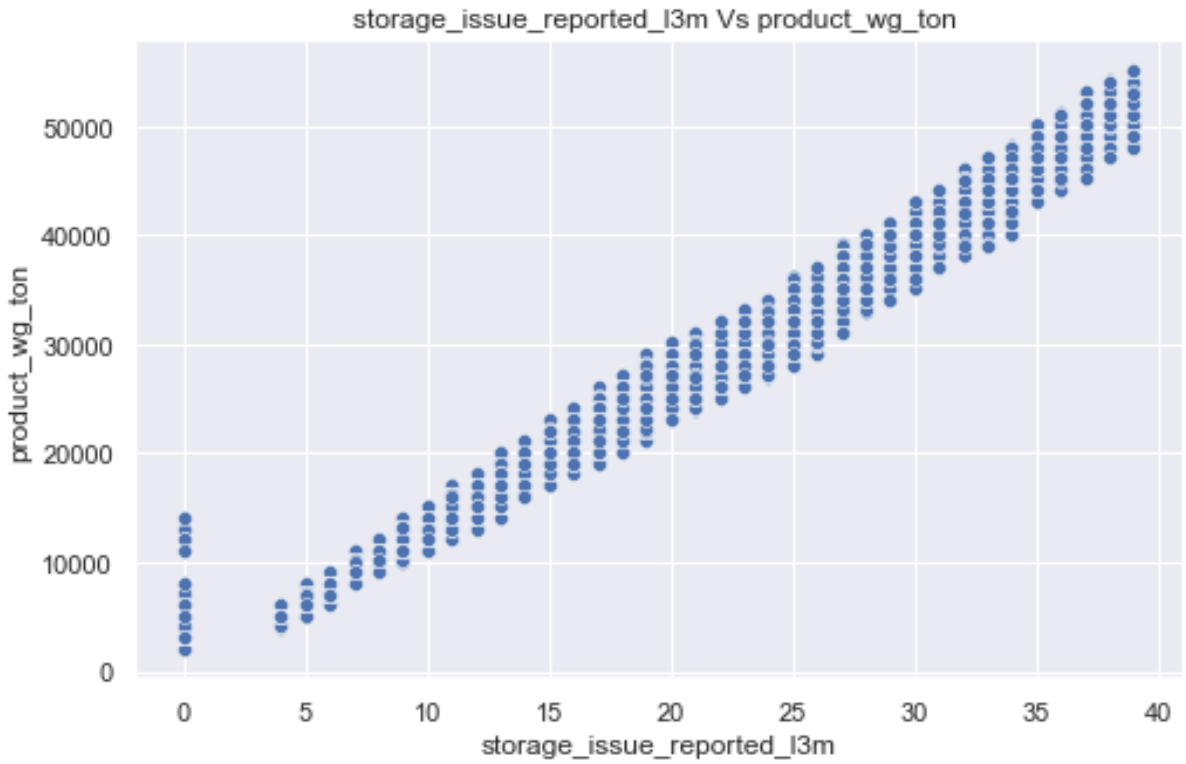
New warehouses are not being set up in East zone.



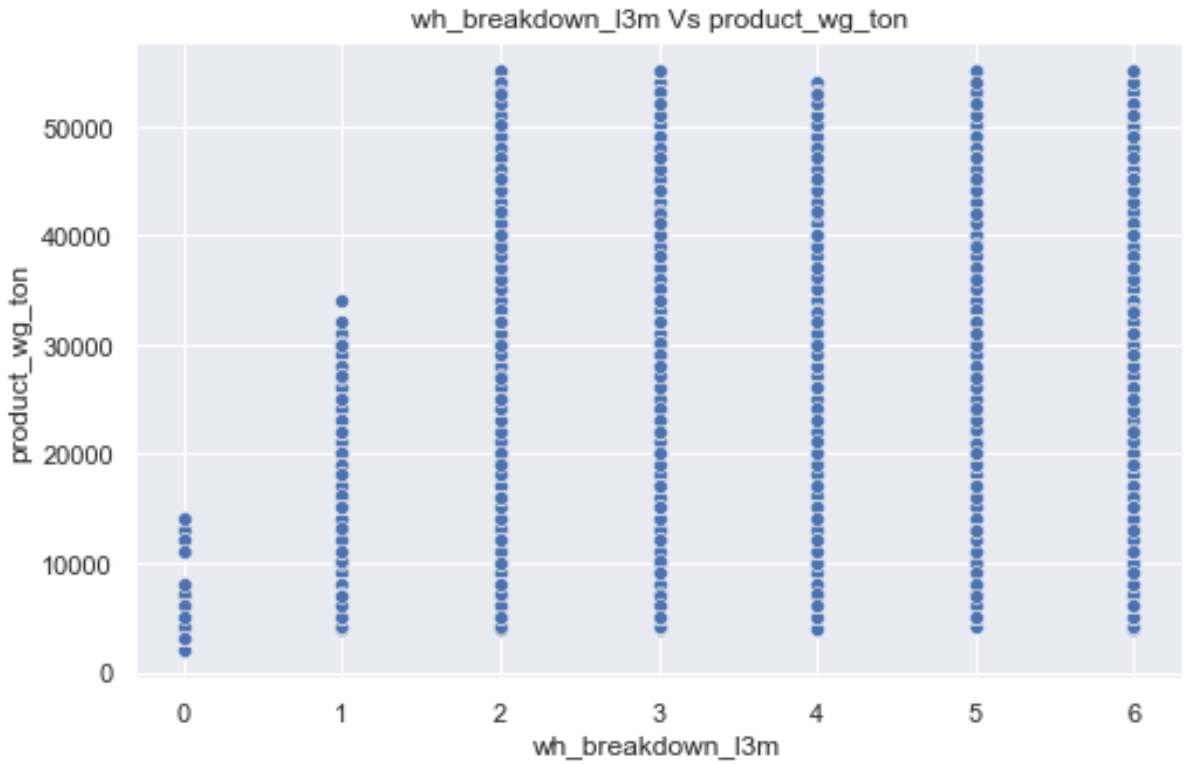
There can be seen a declining trend in product shipment as the number of transport issue increases.



As the number of competitors in the market increases the number of retail shops through which the products being sold are declining.

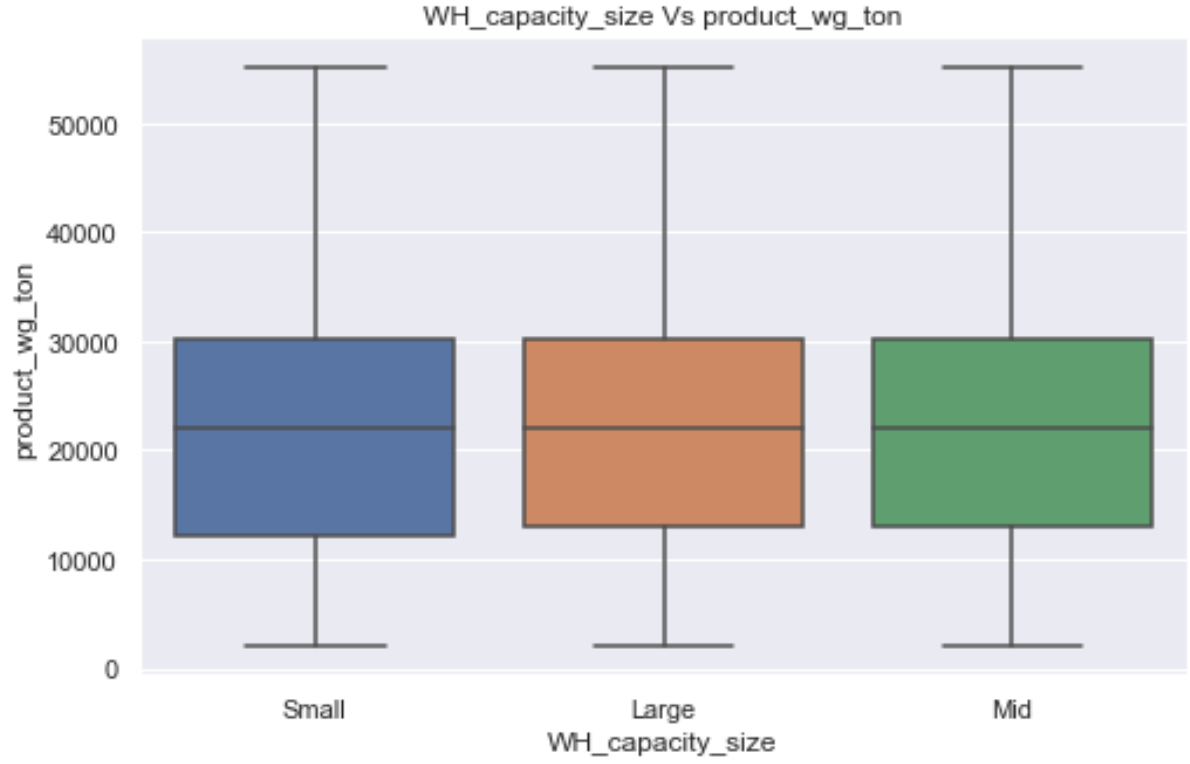


There is a strong linear relationship between the storage issue reported and the product shipped to warehouses.

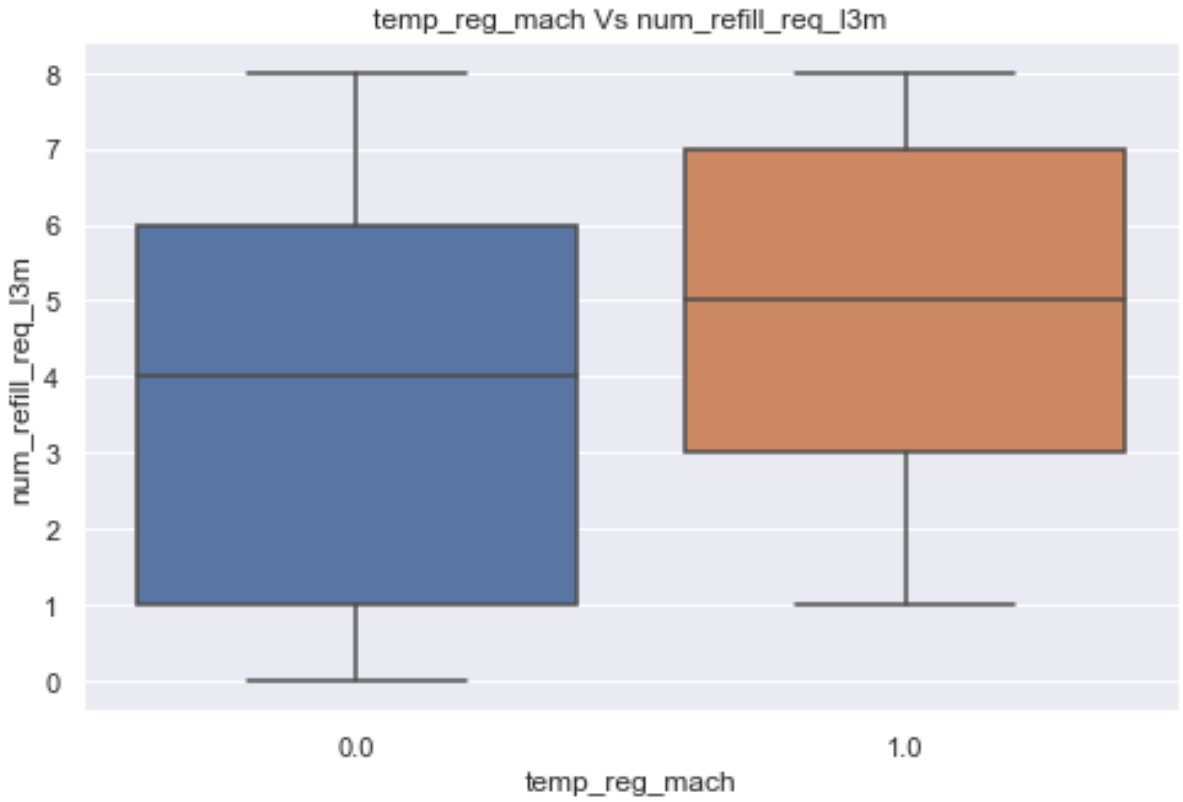


The warehouse breakdowns are frequent as the prduct shipment quantity increases.

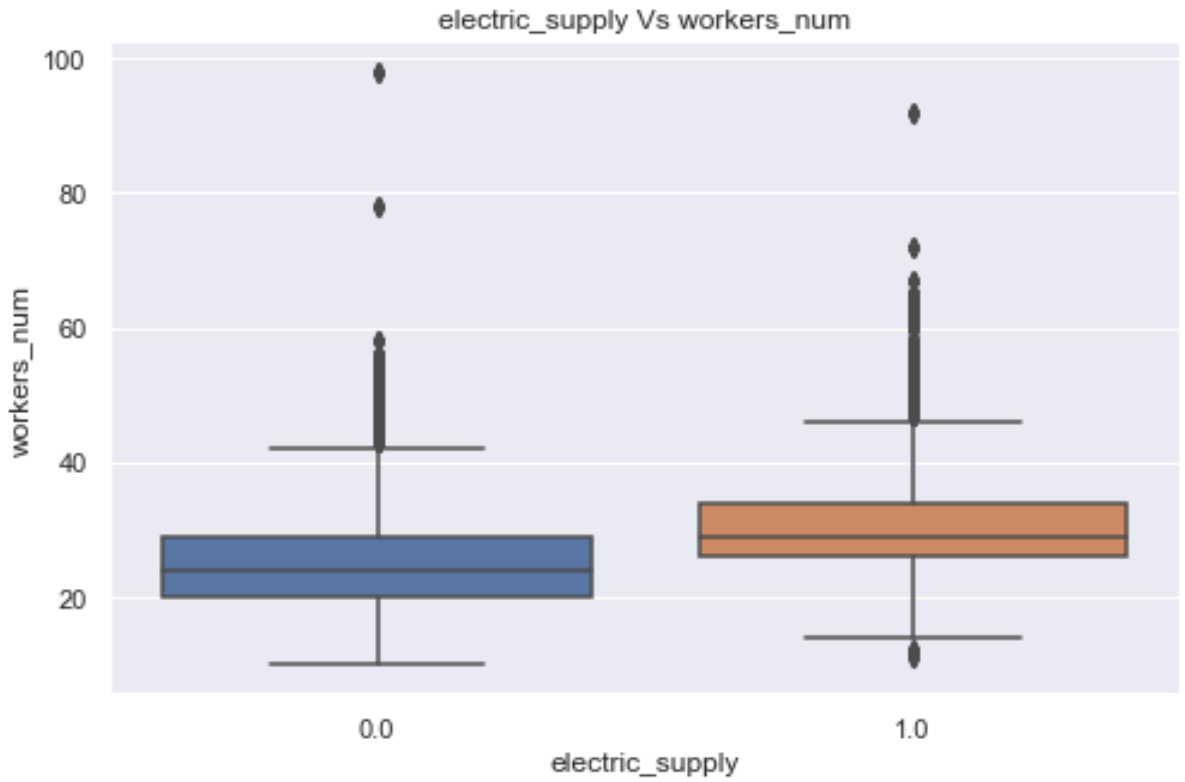
The median age of warehouses in Urban regions are higher than those in rural areas.New warehouses are being established in Rural areas.



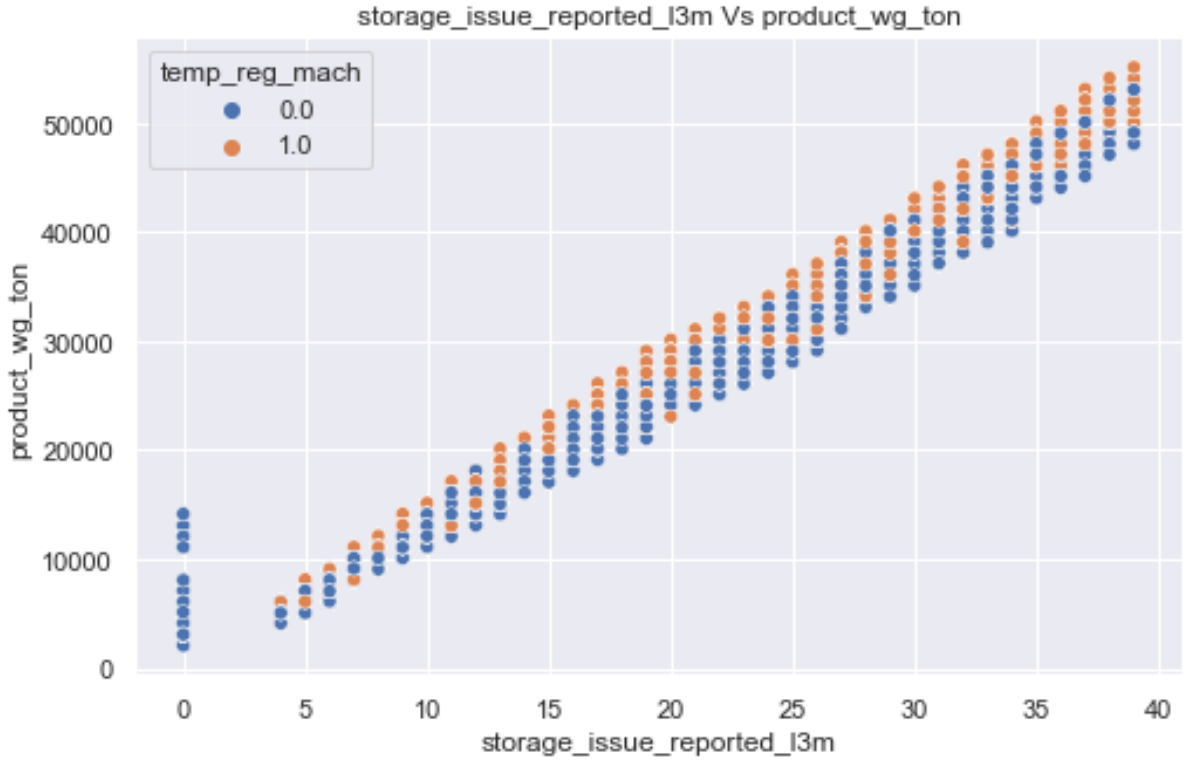
Irrespective of the size of the warehouses the product shipment quantity range is similar.

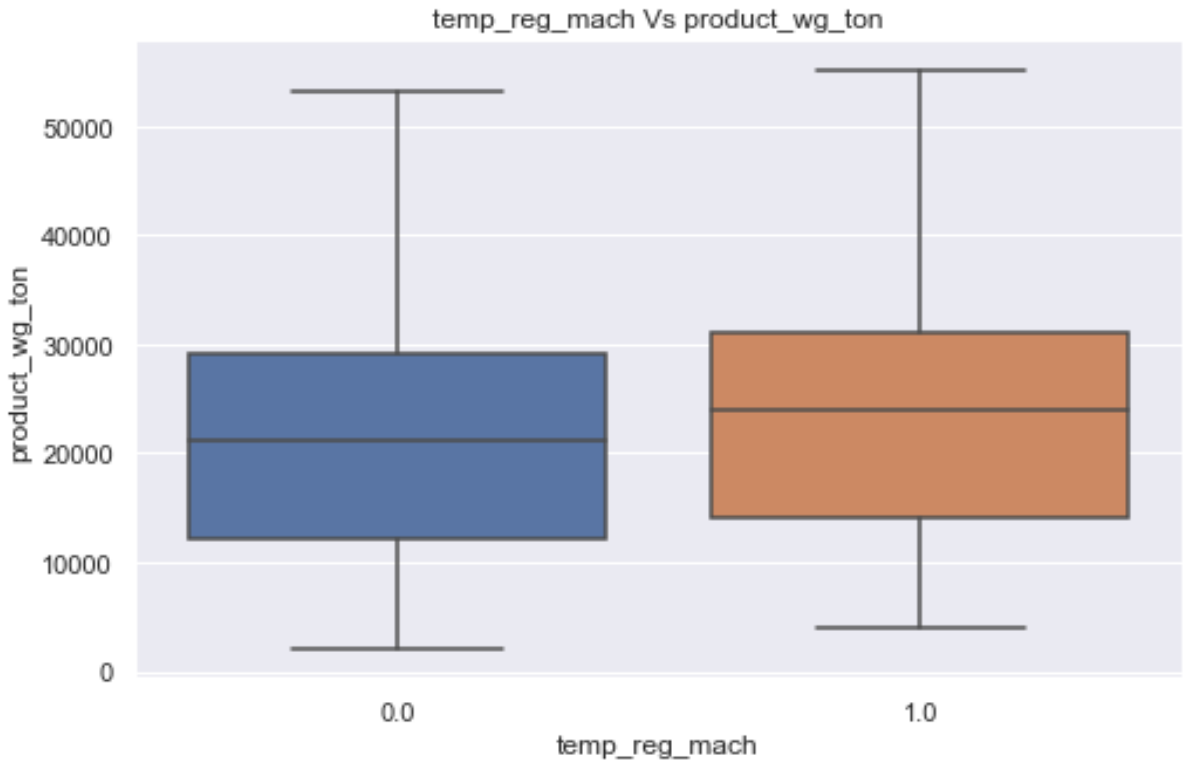


The warehouses with temp_reg_machine indiactor has significantly higher refilling rate than the ones with no indicators installed.



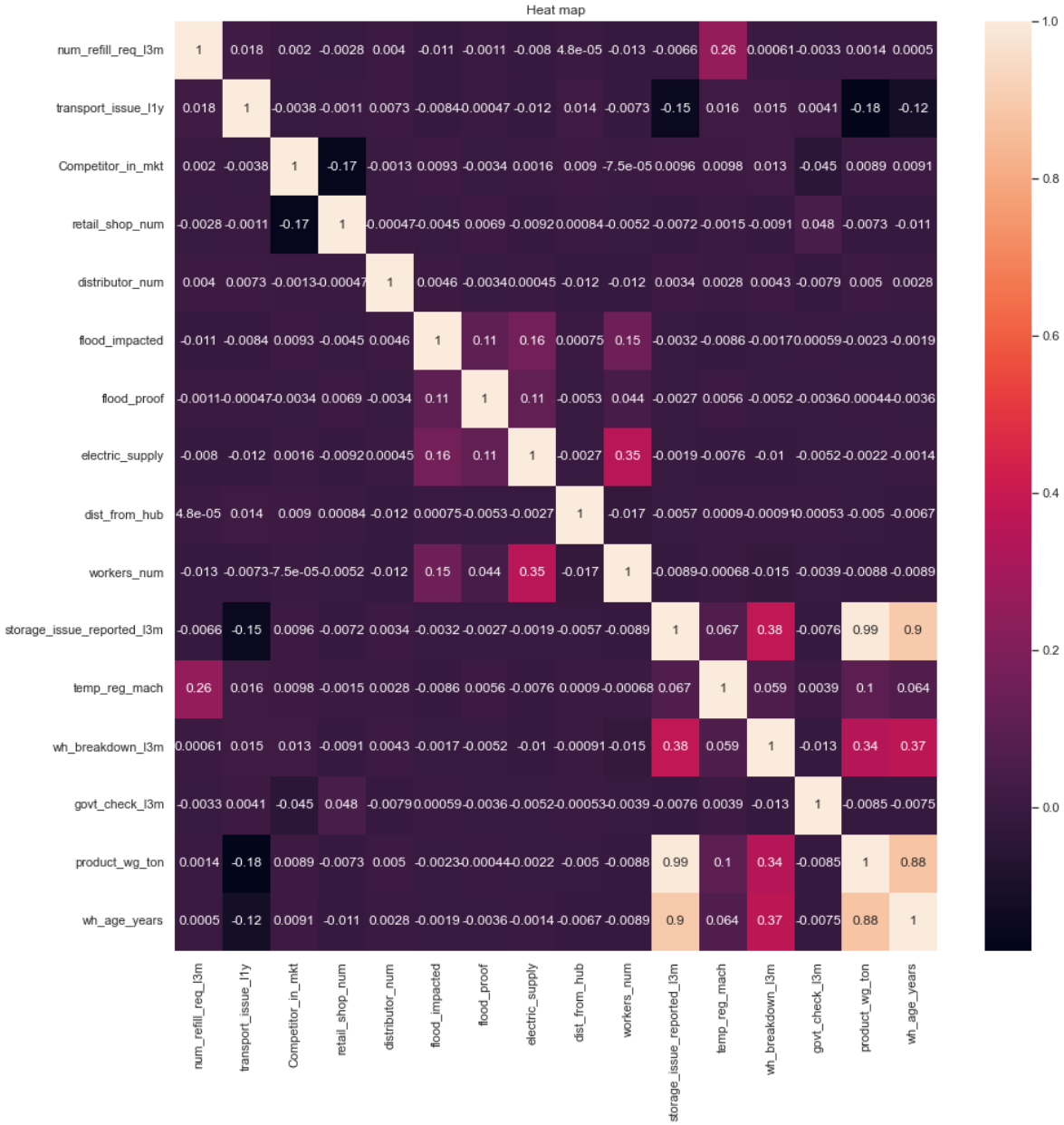
The warehouses with electric supply backup provided employs higher workers than others.





The warehouses with temperature regulating machine indicator received more products

Multivariate Analysis



There is a strong correlation between product_wg_ton and storage_issue_reported_l3m.

There is also strong correlation between product_wg_ton and wh_age_years.

3. Data Cleaning and Pre-processing

- Approach used for identifying and treating missing values and outlier treatment (and why)

Missing Value Treatment

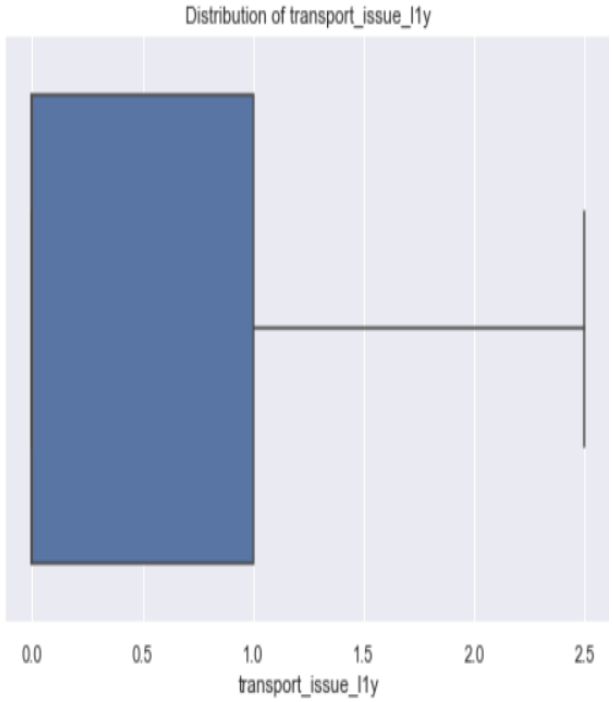
The Variables with missing values are treated as follows

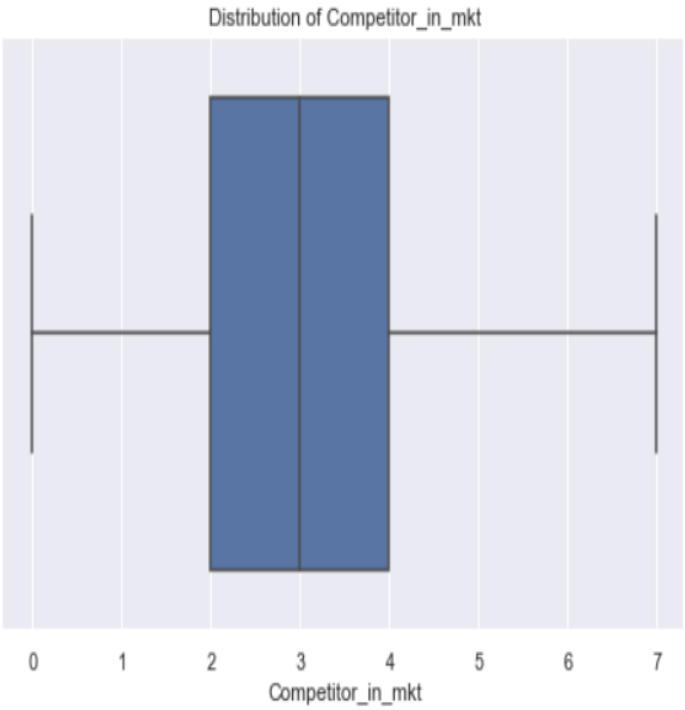
Variables	Missing value Treatment
workers_num	Median
wh_est_year	Feature Engineered into new variable **'Wh_age_years'** & imputed using **KNN imputation**
approved_wh_govt_certificate	Replaced with New category **'application pending'**

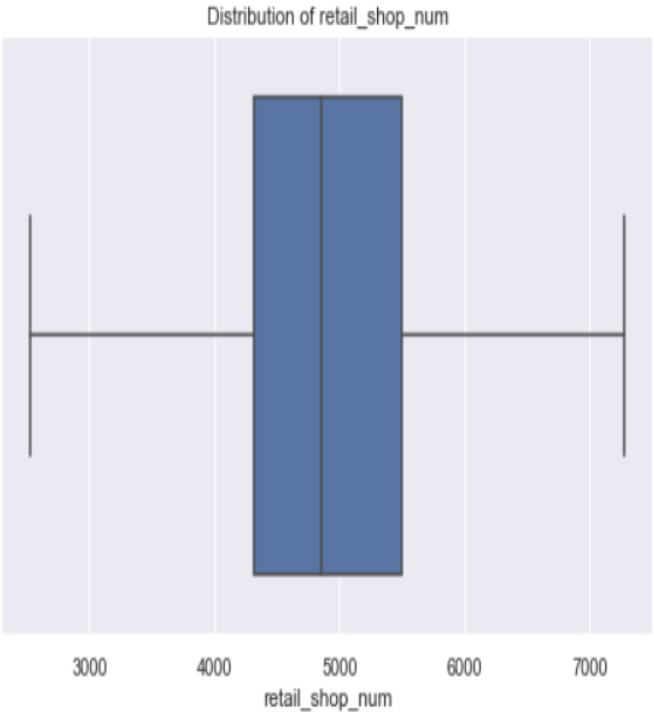
Outlier Treatment

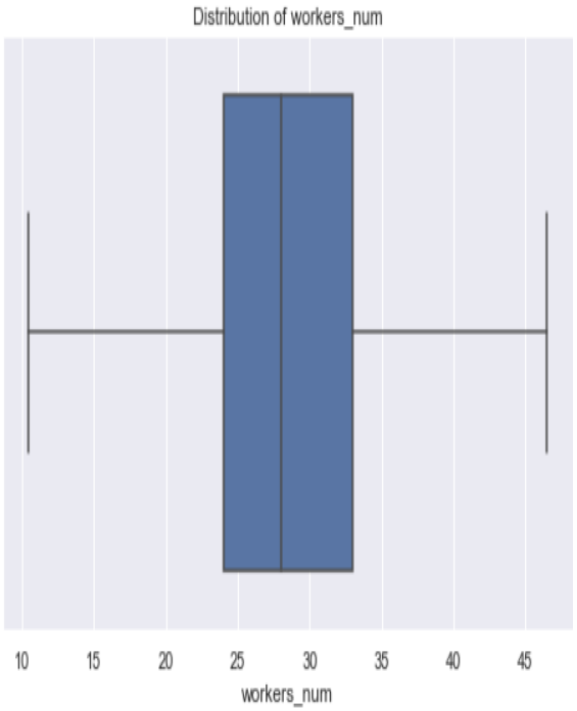
The outliers are treated by relacing them with the nearest whisker values

Outlier Treatment		
Variables	lr	ur
transport_issue_l1y	-1.5	2.5
Competitor_in_mkt	-1	7
workers_num	11	47
retail_shop_num	2532	7280


Distribution of transport_issue_l1y
0.0
0.5
1.0
1.5
2.0
2.5
transport_issue_l1y


Distribution of Competitor_in_mkt
0
1
2
3
4
5
6
7
Competitor_in_mkt





The Above boxplots shows the removal of outliers in the dataset.

Need for variable transformation :

The variable transformations such as feature engineering, categorical data encoging ,scaling are done as the prerequisite measures for model building.

Feature engineering is done for variable **Wh_est_year** it is replaced with a new variable '**wh_age_years** ' by calculating the age based on the current year.

The dataset is split into train & test set

Train Set	70%
Test Set	30%

The categorical variables are encoded using one hot encoding & Label encoding

Variables	Datatype	Transformation
Location_type	object	One Hot Encoding
zone	object	One Hot Encoding
WH_regional_zone	object	One Hot Encoding
wh_owner_type	object	One Hot Encoding

Variables	Datatype	Transformation
WH_capacity_size	object	Label Encoding
approved_wh_govt_certificate	object	Label Encoding

The dataset is then scaled using standard scaler for model building.

SAMPLE DATASET SCALED

	WH_capacity_size	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	...
0	-1.632547	-0.417807	0.374779	-0.980772	-0.317618	-1.146546	-0.329915	4.159520	0.722737	-1.156575	...
1	1.056278	-1.568750	-0.714377	0.803748	1.297843	0.285226	-0.329915	-0.240412	0.722737	0.740827	...
2	-0.288135	-1.185102	-0.714377	0.803748	-0.673514	1.343493	-0.329915	-0.240412	-1.383630	-0.040456	...
3	-0.288135	1.116783	2.008512	-0.980772	1.073989	0.471979	-0.329915	-0.240412	-1.383630	-0.965240	...
4	1.056278	-0.417807	0.374779	-0.980772	-0.225807	-0.026028	3.031081	-0.240412	0.722737	-0.821739	...

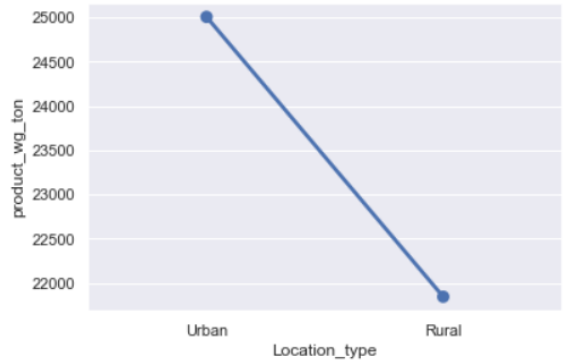
- Variables removed:

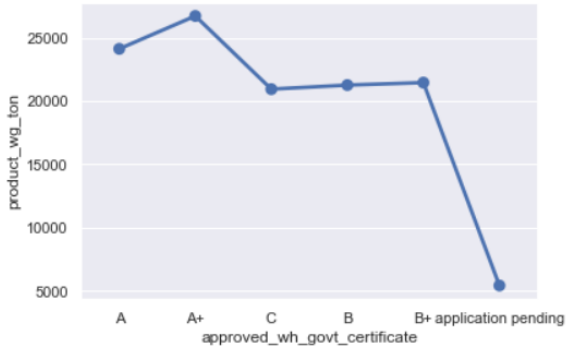
The variables such as **wh _id** and **wh _Manager _id** are dropped from the dataset and not used for modelbuilding

One Way Anova:

One way Anova is used to compare the means of different groups by analyzing the variance.

Oneway Anova	
Variables	P value
Location_Type	3.63E-32
app_gov_cert	0





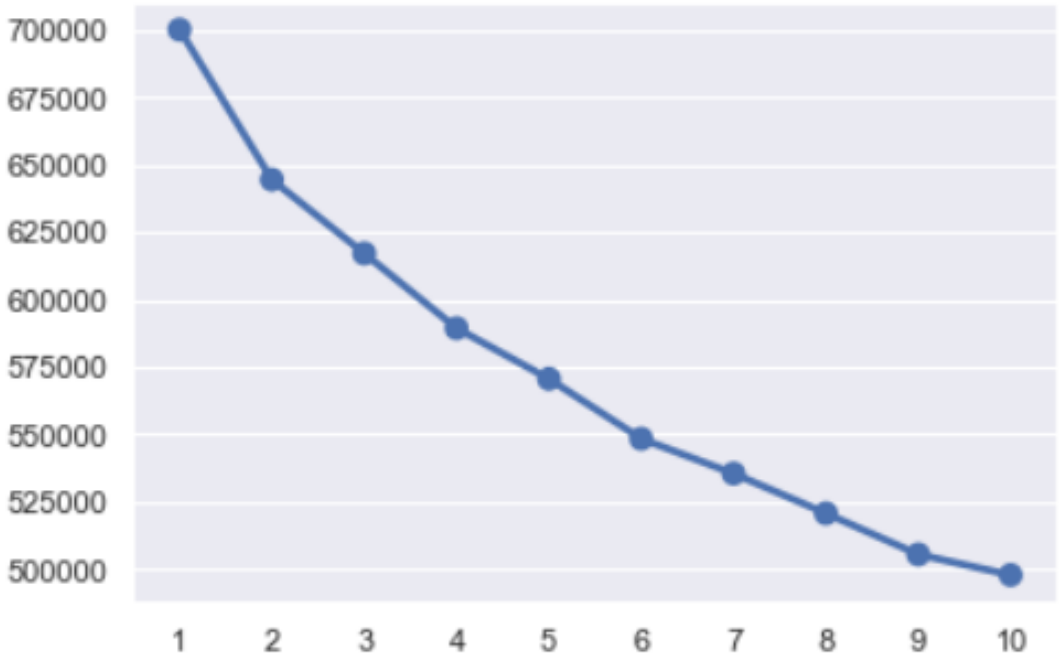
From Anova it is found that the categorical groups within the variables ‘ Location type’ and ‘Approved wh_govt_certificate’ have different means.

Kmeans Clustering:

Kmeans clustering is done on the dataset .Based on the wss plot and silhouette score the dataset is clustered into two clusters ‘cluster 0’ and ‘cluster 1’.

K-MEANS CLUSTERING	WSS
1	700000.00
2	644390.87
3	617029.75
4	589320.17
5	570563.57
6	548364.33
7	535450.55
8	520720.57
9	505476.60
10	497845.18

WSS PLOT

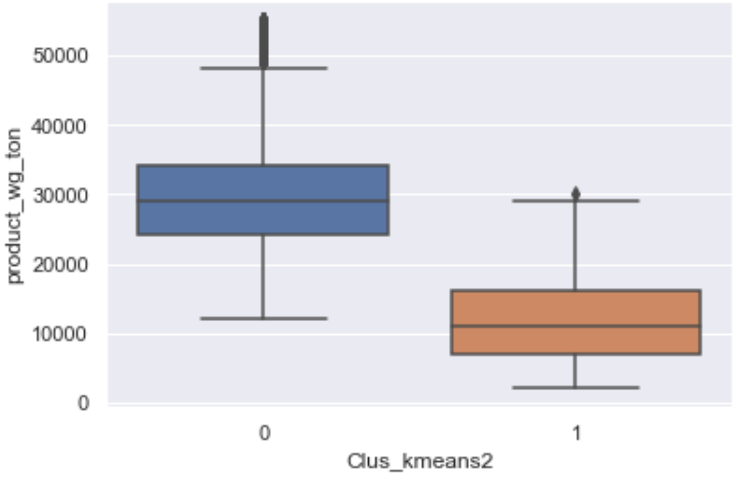


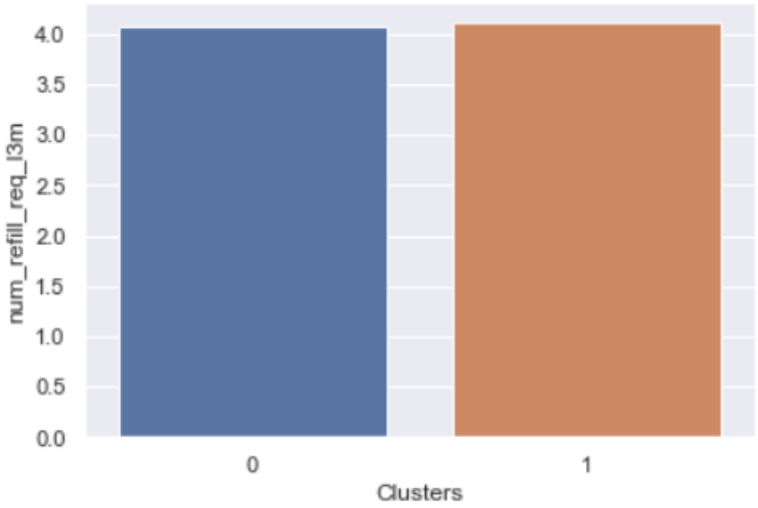
CLUSTER	NO.OF RECORDS
0	14107
1	10893

K-MEANS 2 CLUSTER	SILHOUETTE SCORE	0.076166829

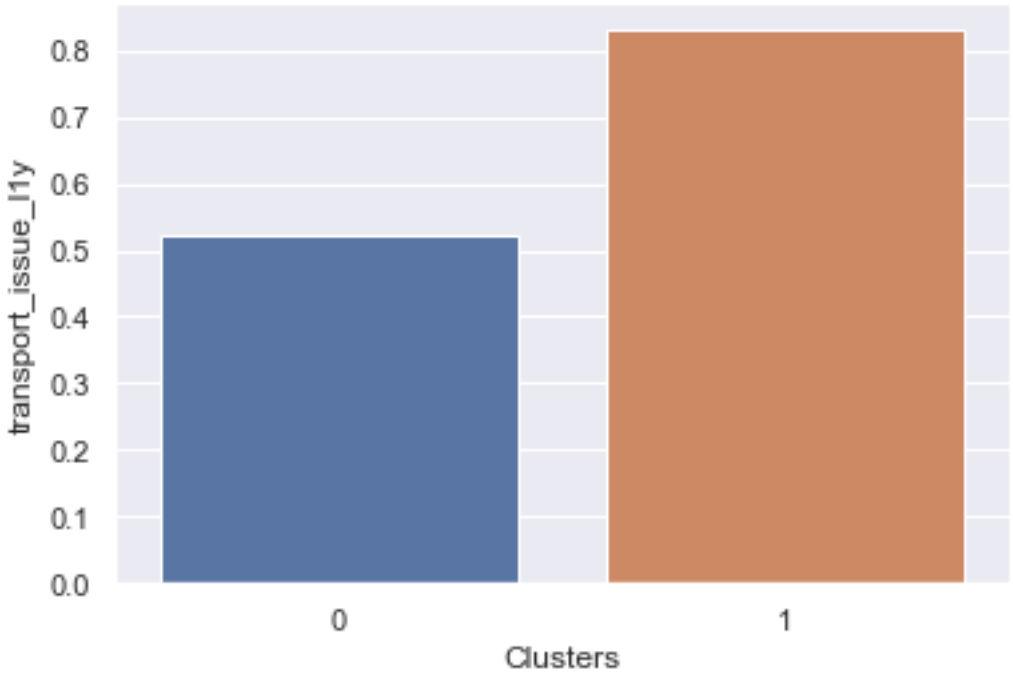
The average values of the variables are obtained for the two clusters and tabulated as follows

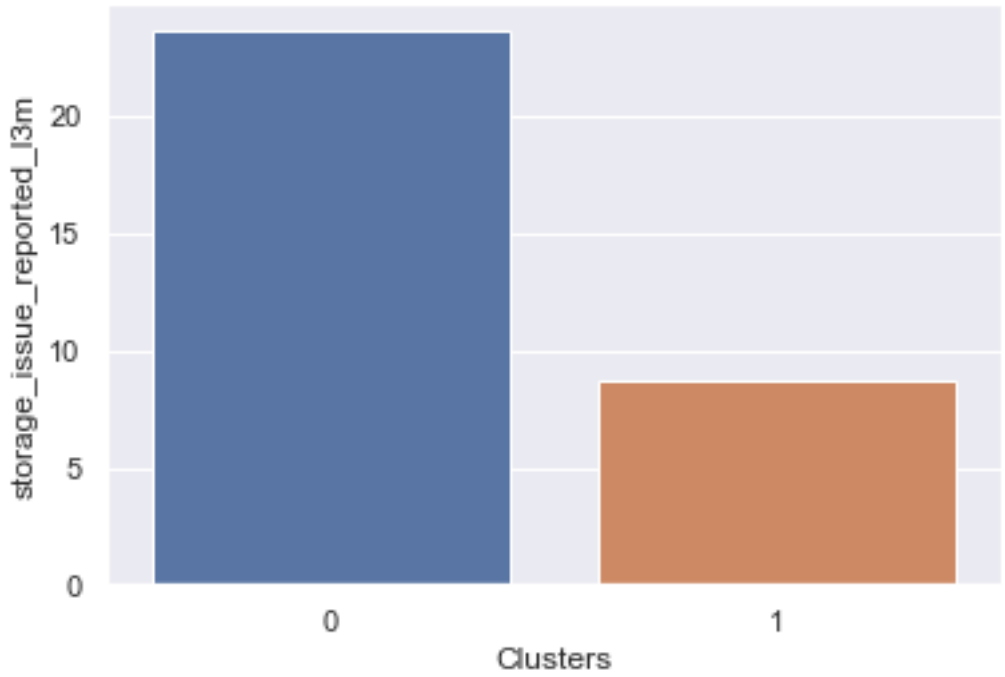
Clus_kmeans2	0	1
num_refill_req_l3m	4.073013	4.109795
transport_issue_l1y	0.521691	0.829707
Competitor_in_mkt	3.116183	3.077206
retail_shop_num	4951.533955	4968.423804
distributor_num	42.439853	42.389975
dist_from_hub	162.948749	164.29955
workers_num	28.661409	28.833746
storage_issue_reported_l3m	23.643936	8.695125
wh_breakdown_l3m	4.057985	2.736161
govt_check_l3m	18.700929	18.956486
product_wg_ton	30270.6024	11524.68879
freq	14107	10893

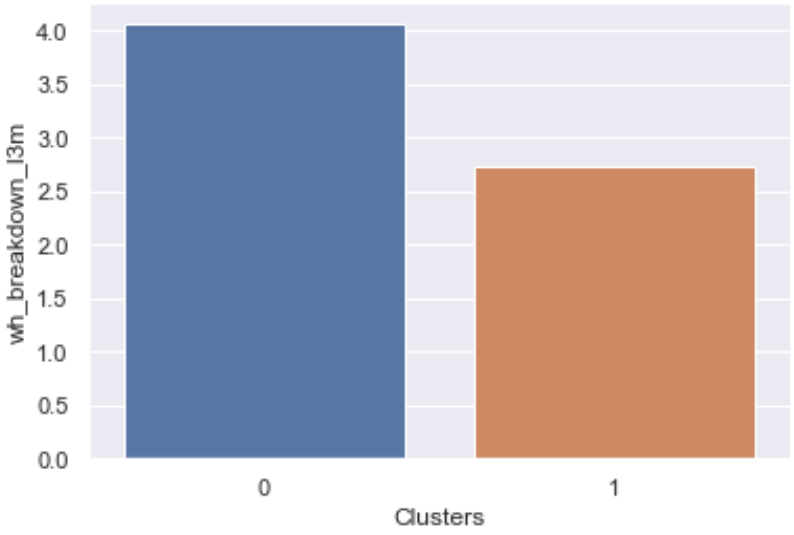


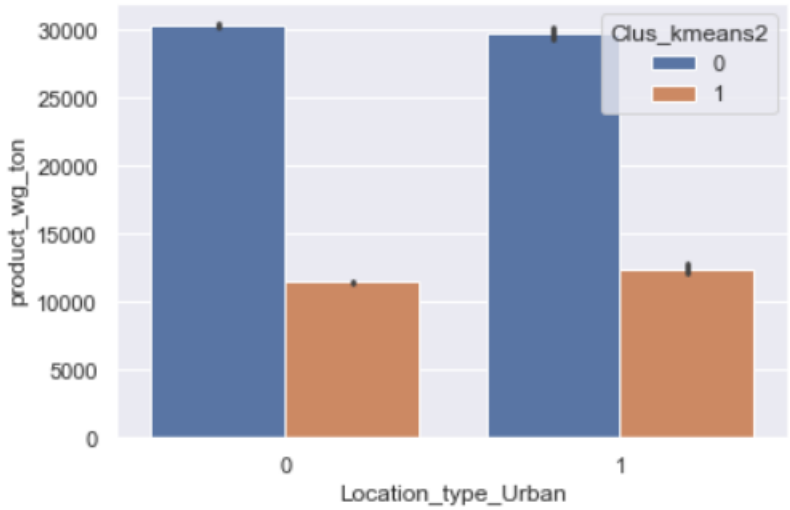


The warehouses in the cluster 0 received significantly higher products than those shipped to cluster 1 warehouses even when No. of times refilling required is similar.

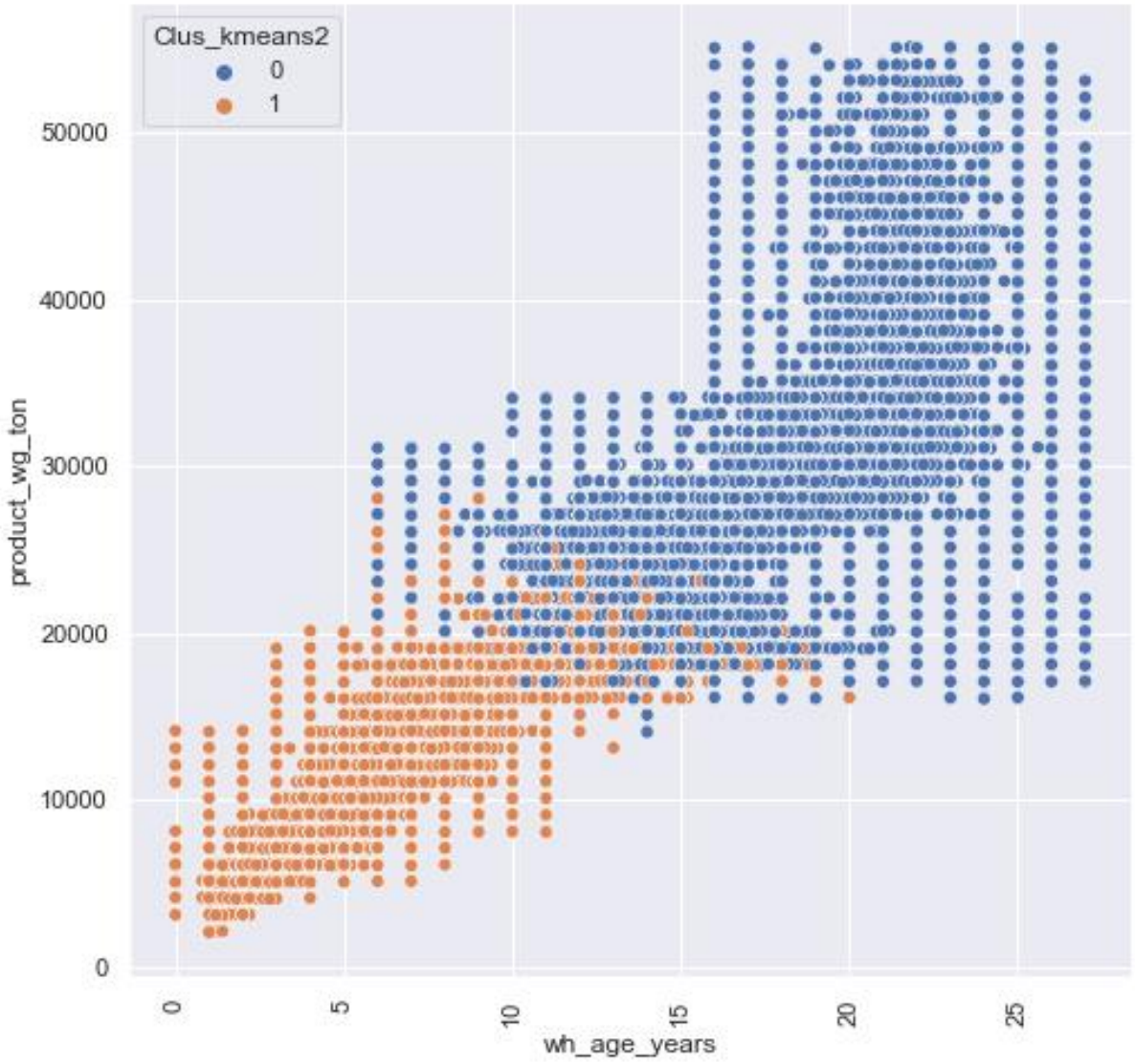

transport_issue_l1y
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
Clusters


storage_issue_reported_l3m
0
5
10
15
20
0
1
Clusters

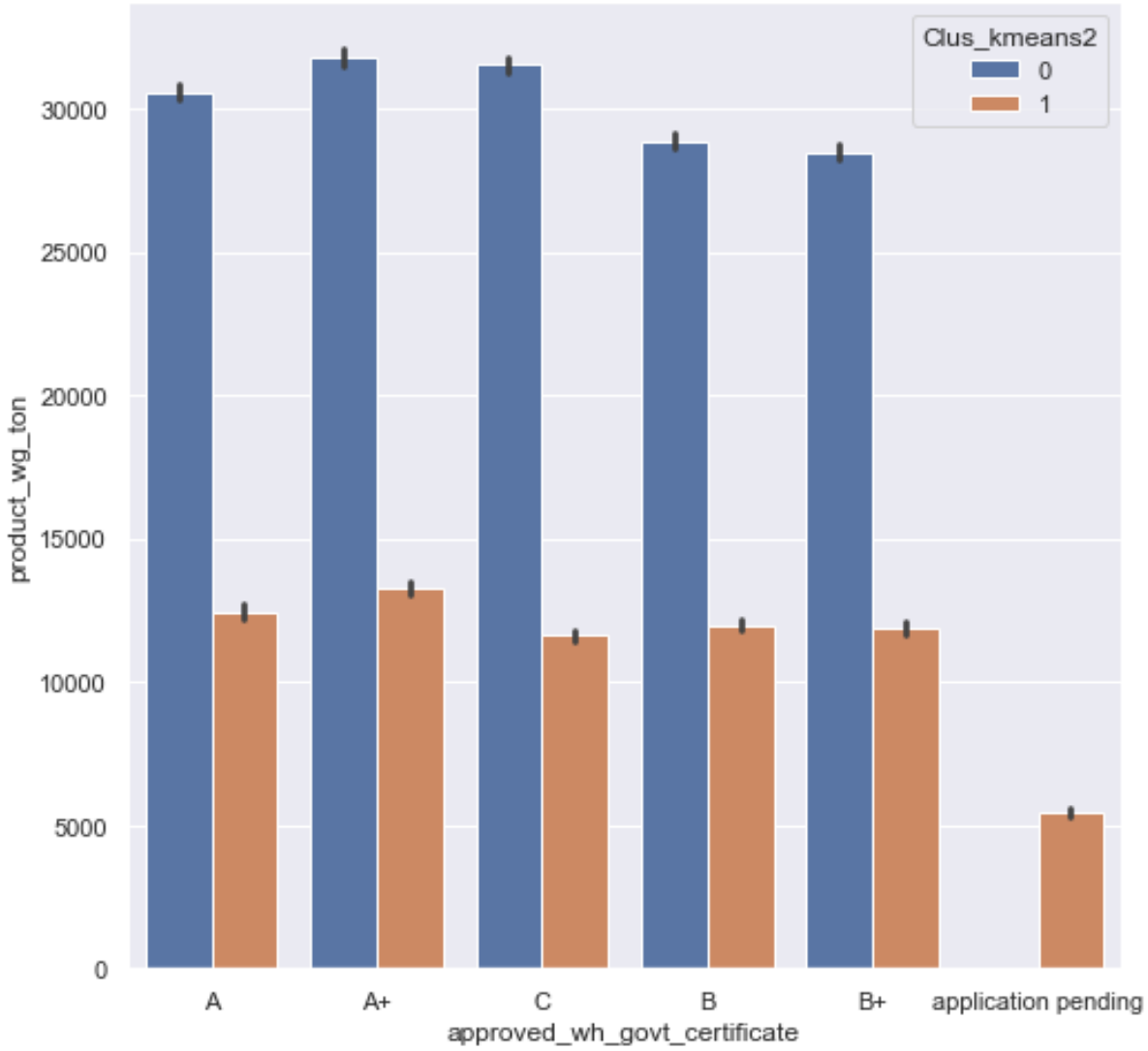




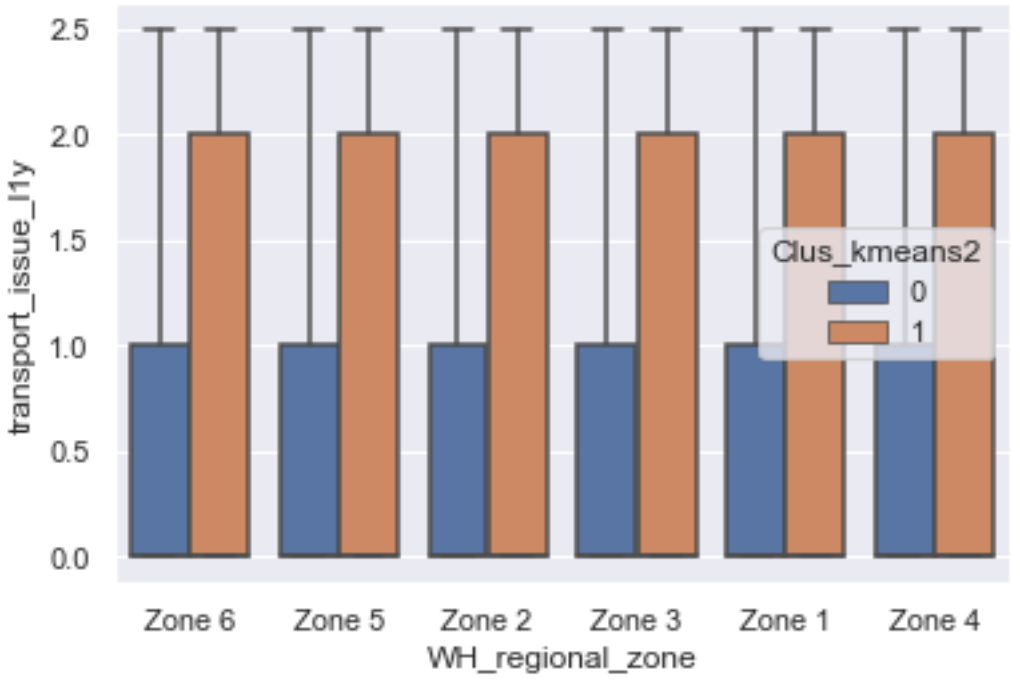
Warehouse breakdown & storage issue reported is higher in cluster 0. Clusters are spread in equal ratio in urban and rural areas



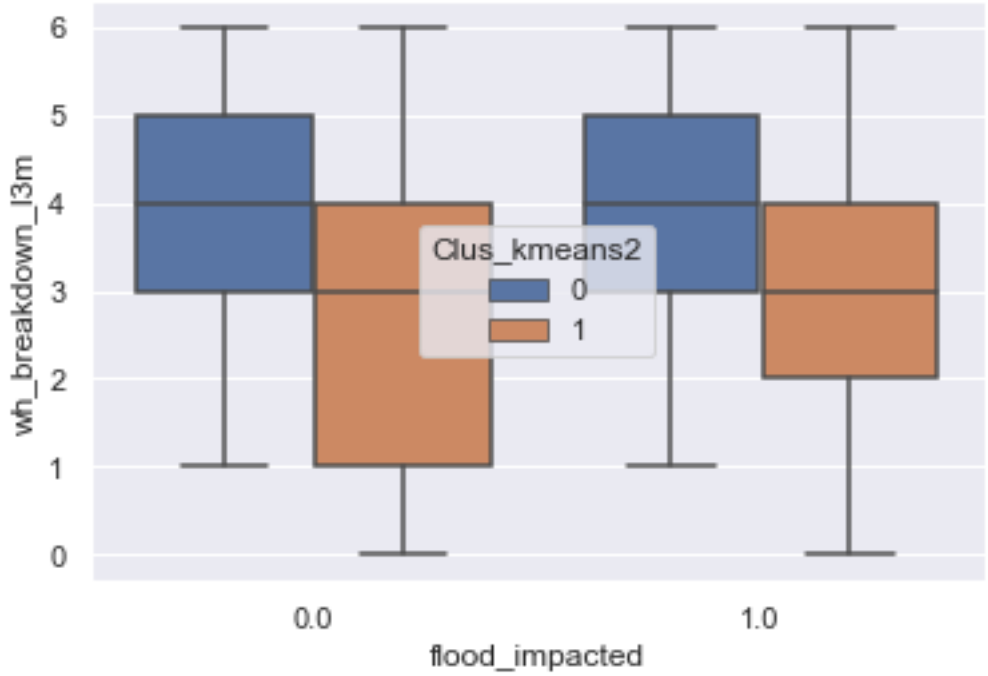
Cluster 0 has higher warehouse age
Cluster 1 has lower warehouse age

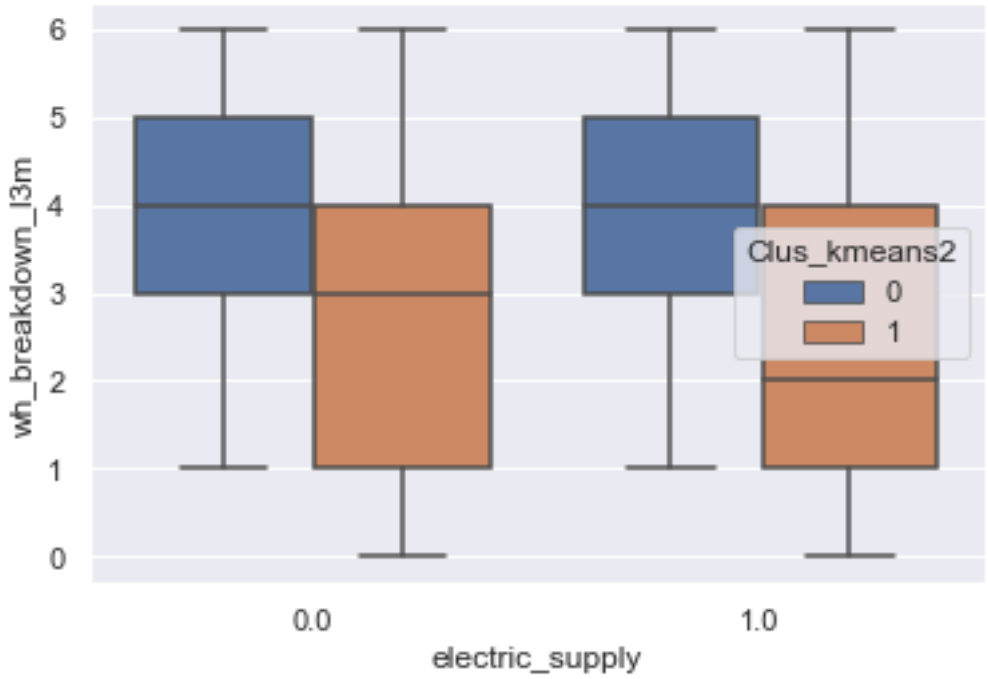


Warehouses in cluster 0 are all certified warehouse

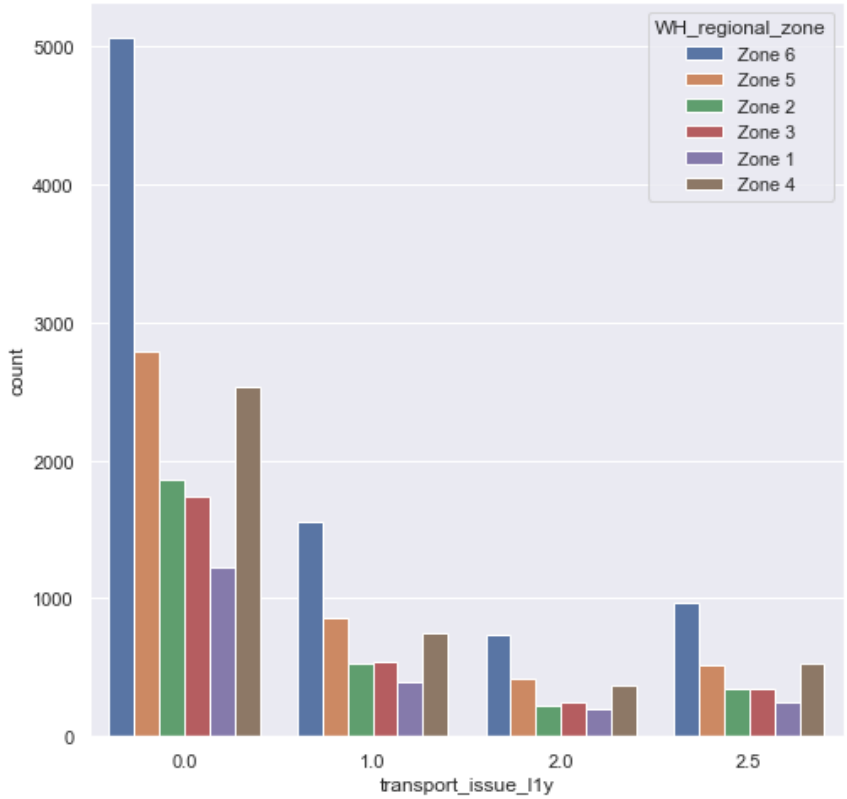


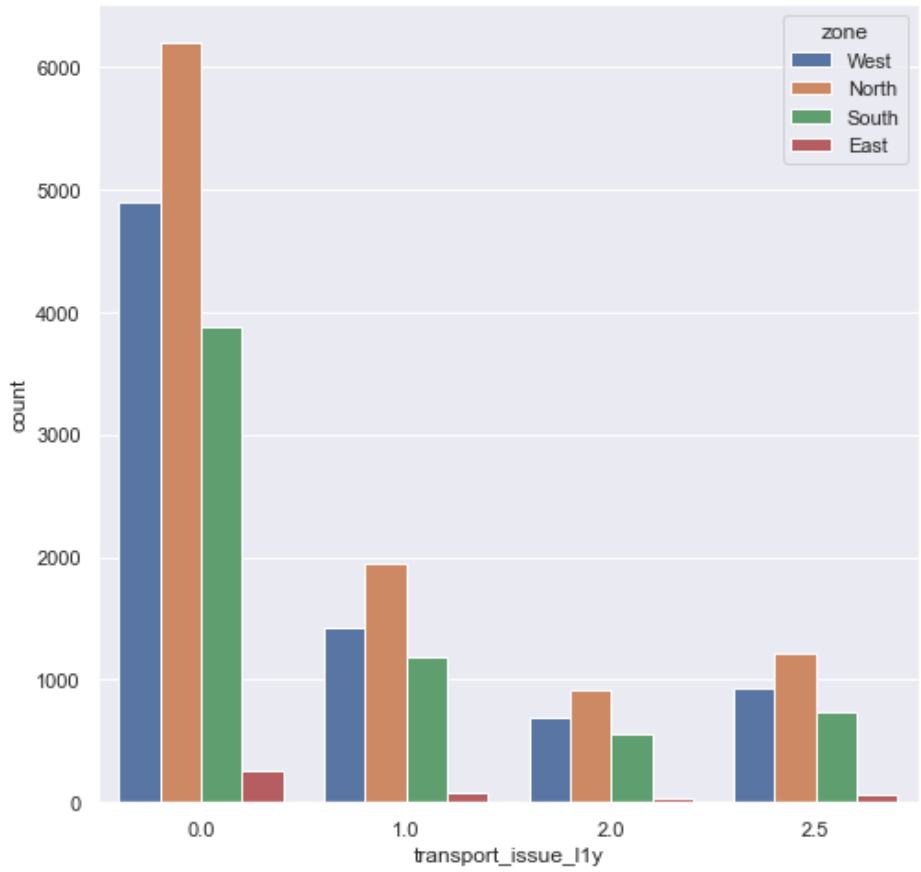
Traffic issues are higher in warehouses of cluster1 across all zones.



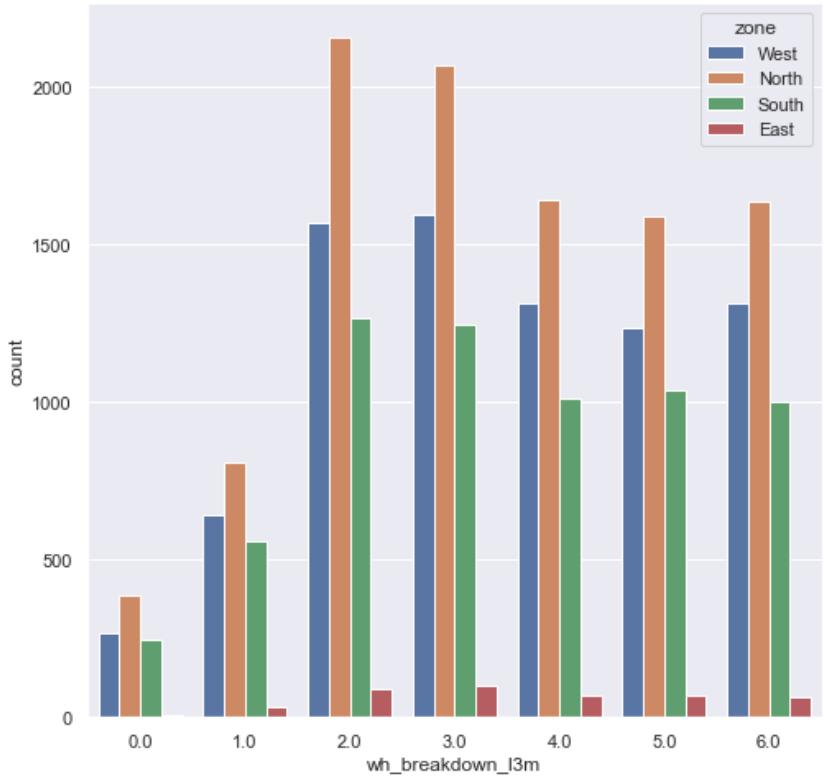


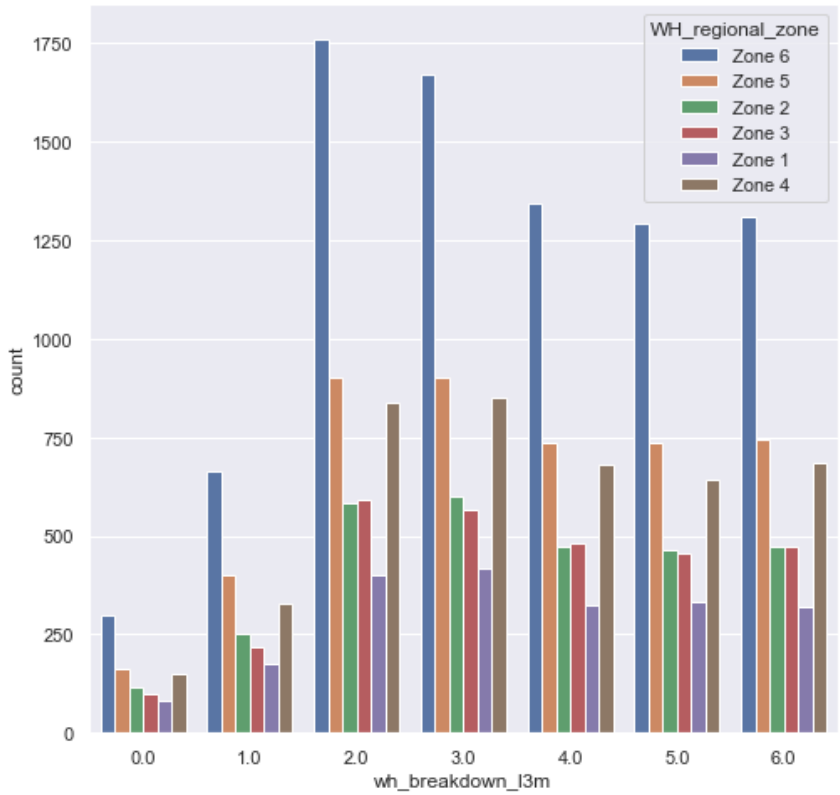
The warehouse breakdowns are similar in cluster 0 between flood impacted and flood unimpacted.
Also it is similar between electricity backup provided & not provided.
Cluster 1 has frequent breakdowns despite flood unimpacted & electricity backup provided.





Transport issues are highly reported in zone 6 among other regional zones and in the North zone.





The warehouse breakdowns are higher in North zone and in zone6 among the regional zones.

North Zone Profile:

	num_refill_req_l3m		product_wg_ton		storage_issue_reported_l3m		transport_issue_l1y		wh_breakdown_l3m	
Clus_kmeans2	0	1	0	1	0	1	0	1	0	1
WH_regional_zone										
Zone 1	1821.0	1508.0	14022598.0	4443533.0	10972.0	3358.0	249.5	333.5	1874.0	1051.0
Zone 2	1982.0	1413.0	14760500.0	4205832.0	11516.0	3132.0	259.5	270.0	1997.0	953.0
Zone 3	2193.0	1735.0	16587375.0	4748360.0	12974.0	3608.0	301.5	317.0	2288.0	1116.0
Zone 4	2673.0	2153.0	20348654.0	5905865.0	15952.0	4425.0	365.0	457.0	2691.0	1344.0
Zone 5	4452.0	3449.0	33276561.0	9616554.0	25957.0	7252.0	602.5	713.0	4444.0	2238.0
Zone 6	10457.0	7978.0	77483018.0	22766973.0	60449.0	17165.0	1317.0	1633.0	10238.0	5377.0

South Zone Profile:

	num_refill_req_l3m		product_wg_ton		storage_issue_reported_l3m		transport_issue_l1y		wh_breakdown_l3m	
Clus_kmeans2	0	1	0	1	0	1	0	1	0	1
WH_regional_zone										
Zone 1	1362.0	1298.0	10890973.0	3791893.0	8537.0	2791.0	200.0	262.5	1484.0	849.0
Zone 2	3465.0	2467.0	25607555.0	6860344.0	19972.0	5196.0	417.0	482.0	3462.0	1597.0
Zone 3	1884.0	1768.0	14022234.0	4787885.0	10960.0	3591.0	247.0	362.0	1886.0	1145.0
Zone 4	1973.0	1651.0	14665843.0	4564827.0	11501.0	3461.0	264.5	324.5	1979.0	1084.0
Zone 5	2457.0	2039.0	18270722.0	5842975.0	14293.0	4445.0	303.5	359.5	2483.0	1390.0
Zone 6	3103.0	2460.0	23089734.0	7145916.0	18034.0	5385.0	408.5	528.0	3089.0	1607.0

East Zone Profile:

	num_refill_req_l3m		product_wg_ton		storage_issue_reported_l3m		transport_issue_l1y		wh_breakdown_l3m	
Clus_kmeans2	0	1	0	1	0	1	0	1	0	1
WH_regional_zone										
Zone 1	92.0	77.0	623285.0	249053.0	491.0	195.0	12.5	16.5	96.0	52.0
Zone 3	285.0	165.0	2019791.0	506893.0	1571.0	391.0	31.0	42.5	268.0	139.0
Zone 4	342.0	246.0	2594436.0	711735.0	2016.0	550.0	40.0	62.5	312.0	185.0
Zone 5	165.0	115.0	1416385.0	351689.0	1105.0	270.0	19.5	26.5	169.0	99.0
Zone 6	146.0	114.0	1008628.0	265608.0	795.0	210.0	20.5	18.0	135.0	60.0

West Zone Profile:

	num_refill_req_l3m		product_wg_ton		storage_issue_reported_l3m		transport_issue_l1y		wh_breakdown_l3m	
Clus_kmeans2	0	1	0	1	0	1	0	1	0	1
WH_regional_zone										
Zone 1	1085.0	879.0	8141270.0	2496927.0	6341.0	1870.0	122.0	194.5	1104.0	596.0
Zone 2	1641.0	1113.0	12260539.0	2885998.0	9548.0	2153.0	177.5	234.5	1548.0	731.0
Zone 3	2341.0	1531.0	16451602.0	4166090.0	12852.0	3158.0	263.5	338.5	2235.0	1060.0
Zone 4	4631.0	3565.0	34025625.0	9779044.0	26541.0	7413.0	568.5	714.0	4597.0	2422.0
Zone 5	3290.0	2764.0	24821238.0	7421489.0	19359.0	5571.0	436.5	534.5	3371.0	1814.0
Zone 6	5618.0	4280.0	40638822.0	12022952.0	31809.0	9126.0	732.5	814.0	5496.0	2896.0

4. Model building

- Clear on why was a particular model(s) chosen.

To predict the optimum weight to be shipped to each warehouse the following regression models are used.

1.Decision Tree

- Decision Trees are popular choice for regression problems as they are simple & easy to interpret.

- Captures non-linear relationships between the input and target variables.

- They split the input variables into smaller regions, based on the values of the input variables, until each region contains a relatively homogeneous set of data points with similar target variable values.

- The prediction for a new data point is then simply the average of the target variable values in the region that the point falls into.

2. Random Forest

- Random Forest is a popular choice for regression problems because it combines the
- simplicity and interpretability of decision trees with the power of ensemble methods.

- In a Random Forest model, multiple decision trees are trained on different subsets of the training data and with different subsets of the input variables.

- Each tree in the forest makes an independent prediction, and the final prediction is an average or a weighted average of the predictions from all the trees.

- This approach helps to reduce Overfitting & improve performance of model.

Advantages:

- It can capture complex non-linear relationships between the input variables and the target variable.

- It can handle high-dimensional input data and interactions between variables more effectively than a single decision tree.

- Provides a measure of feature importance, which can be used to identify the most elevant input variables for the target variable.

3. Linear regression

- It is a statistical method that models the linear relationship between the input variables and the target variable.

- The model assumes that the target variable is a linear combination of the input variables, plus an error term that accounts for the variability in the data.

- It can be used to estimate the magnitude and direction of the effect of each input variable on the target variable.

Assumptions:

- It is assumed that relationship between the input variables and the target variable is linear.

- The errors are normally distributed and have constant variance, which may not hold for all datasets.

4.Artificial Neural Network(ANNs)

- ANNs are a popular choice for regression problems. ANNs consist of layers of interconnected nodes that process the input data and generate an output.

- Each node applies a mathematical function to the input data and passes the result to the next layer of nodes. The final layer produces the output prediction.

Advantages:

- Handle large and complex datasets with many input variables.

- Automatically extract relevant features from the input data, reducing the need for feature engineering.

- They can learn from noisy or incomplete data,

- They can be easily adapted to new data and modified to address specific problems

- Effort to improve model performance.

The model performance can be enhanced by adopting various measures such as Pruning,Ensemble techniques,Feature selection,Feature Engineering,Outlier Treatment,regularization,Hyper parameter tuning etc.

Hyper Prameters Tuning:

The models are built and then a Gridsearch Cross validation is performed to determine the best parameters for tuning the model to improve the performance.

Based on the result of Gridsearch CV the models are tuned by introducing some hyper parameters as follows

MODEL	BEFORE TUNING	AFTER TUNING
Decision Tree	Random state=123	Max_depth= 10,min_samples_leaf=3,min_samples_split= 40
Random Forest regressor	Random state=123	Max_depth= 20, Max_features = 15, min_samples_leaf = 2,min_samples_split = 6,n_estimators = 450
Artificial Neural Networks	Hidden layers = 500, Random state = 123, Max iterations = 1000	Activation='relu', Hidden_layer_sizes = 400, Solver = 'adam'

5. Model validation

How was the model validated? Just accuracy, or anything else too?

Model Evaluation:

Before Hyper parameter Tuning:

MODEL	Train RMSE	Test RMSE	Training Score	Test Score	Training MAPE	Test MAPE
Decision Tree	0	1246.511	1	0.988434	0	5.013337
Random Forest regressor	342.35973	918.3133	0.999131	0.993723	1.542264	4.104171
Linear Regression	1740.1091	1785.863	0.977554	0.97626	8.901757	9.11944
Artificial Neural Networks	983.51515	1105.519	0.99283	0.990902	4.758988	5.369675

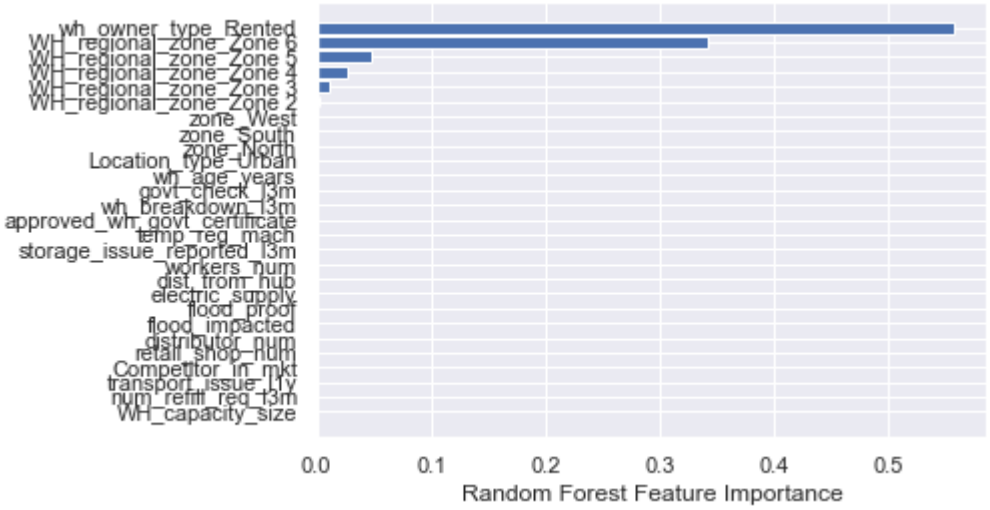
After Hyper parameter Tuning:

MODEL	Train RMSE	Test RMSE	Training Score	Test Score	Training MAPE	Test MAPE
Decision Tree	835.81596	940.6142	0.994822	0.993414	3.718245	4.179332
Random Forest regressor	532.57756	920.843	0.997897	0.993688	2.407422	4.151417
Linear Regression	1740.1091	1785.863	0.977554	0.97626	8.901757	9.11944
Artificial Neural Networks	1043.859	1144.488	0.991923	0.99025	5.138446	5.638679

- All models performs well and there is no overfitting
- The Random Forest regressor has the highest accuracy – 0.9937
- The Random Forest regressor has the least Mean Absolute Perccentage Error(MAPE) – 4.1
- Thus the Random Forest regressor is chosen as best model for prediction.

Feature Importance:

The important features which plays major role in model building to predict the target variables are listed



6. Final interpretation / recommendation

- Detailed recommendations for the management/client based on the analysis done.

Recommendations:

Cluster 0

- Warehouses in Cluster 0 are older and the warehouse breakdowns are higher .Measures to be taken to reduce the warehouse breakdowns.
- Storage issues reported are also higher in these warehouses with higher shipment.
- Despite being supplied products in higher quantity ,the number of refilling required is similar to cluster 1, indicating the fact that these warehouses have higher demand.
- The steps to be taken to equip such warehouses with proper storage facilities so that warehouse breakdowns can be reduced.

Cluster 1

- Despite less shipment quantity , the refilling done is low showing poor movement of stock hence supply to warehouses in cluster 1 can be reduced .
- Traffic issues are higher in these cluster ,it is advisable to inquire the issue & implement corrective measures such as changing mode of transport, equipping the logistic with surveillance devices.

North Zone

- Zone 6 has higher product shipment and higher refilling requirements indicating higher demand. Their supply can be increased further to cater the demand.

- The storage issue,Transport issue and warehouse breakdowns are also higher in zone 6.

- Zone 1 has low shipment and low demand.

South Zone

- Zone 2 has higher product shipment and higher refilling requirements indicating higher demand. Their supply can be increased further to cater the demand.

- The storage issue,Transport issue and warehouse breakdowns are also higher in zone 2.

- Zone 1 has low shipment and low demand.

East Zone

- Zone 4 has higher product shipment and higher refilling requirements indicating higher demand. Their supply can be increased further to cater the demand.

- The storage issue,Transport issue and warehouse breakdowns are also higher in zone 4.

- Zone 1 has low shipment and low demand.

West Zone

- Zone 6 has higher product shipment and higher refilling requirements indicating higher demand. Their supply can be increased further to cater the demand.

- The storage issue,Transport issue and warehouse breakdowns are also higher in zone 6.

- Zone 1 has low shipment and low demand

Supply can be reduced for warehouses with low demand.

Marketing Recommendations

- **Cluster 0** – Marketing campaigns can be aimed for retention & also for acquiring new customers

- **Cluster 1**- Campaigns should focus on creating brand awareness