Great Learning

Capstone Report

Supply Chain Analysis

Submitted by:

N. Aishwarya

PGP-DSBA

December 2022

Business Report Outline

List of Figures

List of Tables

List of Formulas

1. Introduction to the Business Problem

Problem Statement:

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

The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time to the warehouse.

Purpose of Study:

The company is incurring inventory loss due to inadequate demand-supply management of its product, instant noodles. The management needs to optimize the supply quantity across all warehouses in the country.

Hence the objective of this project is to build a model using historical data, to determine an optimum product weight to be shipped each time to the warehouse. With multiple options available to analyse data, it is challenging to decide the appropriate machine learning model to use since the performance of the model vary on the parameters available in the data. This project aims to compare different popular machine learning classifiers, and measure their performance to find out which machine learning model performs better.

Approach:

Data exploration will be carried out using the Python, through univariate and bivariate analysis. The presence of outliers interferes with the correct modelling and interpretation of the data. Thus, the outliers (extreme values) are identified and replaced.

Since the dataset used is related to supply chain, important parameters are identified and the machine learning models are trained with the dataset for detection of optimum weight. The study helps to analyse/optimize the supply quantity at each warehouse in the country & thereby determining the advertising strategies & campaigns for specific pockets.

In this problem, “PRODUCT_WG_TON is the target variable.

Data dictionary as described below:

S. No	Field Name	Description
1	Ware_house_ID	Product warehouse ID
2	WH_Manager_ID	Employee ID of warehouse manager
3	Location_type	Location of warehouse like in city or village
4	WH_capacity_size	Storage capacity size of the warehouse
5	zone	Zone of the warehouse

6	WH_regional_zone	Regional zone of the warehouse under each zone
7	num_refill_req_l3m	Number of times refilling has been done in last 3 months
8	transport_issue_l1y	Any transport issue like accident or goods stolen reported in last one year
9	Competitor_in_mkt	Number of instant noodles competitor in the market
10	retail_shop_num	Number of retails shop who sell the product under the warehouse area
11	wh_owner_type	Company is owning the warehouse or they have got the warehouse on rent
12	distributor_num	Number of distributers works in between warehouse and retail shops
13	flood_impacted	Warehouse is in the Flood impacted area indicator
14	flood_proof	Warehouse is flood proof indicators. Like storage is at some height not directly on the ground
15	electric_supply	Warehouse have electric back up like generator, so they can run the warehouse in load shedding
16	dist_from_hub	Distance between warehouse to the production hub in Kms
17	workers_num	Number of workers working in the warehouse
18	wh_est_year	Warehouse established year
19	storage_issue_reported_l3m	Warehouse reported storage issue to corporate office in last 3 months. Like rat, fungus because of moisture etc.
20	temp_reg_mach	Warehouse have temperature regulating machine indicator
21	approved_wh_govt_certificate	What kind of standard certificate has been issued to the warehouse from government regulatory body
22	wh_breakdown_l3m	Number of time warehouse face a breakdown in last 3 months. Like strike from worker, flood, or electrical failure
23	govt_check_l3m	Number of time government Officers have been visited the warehouse to check the quality and expire of stored food in last 3 months
24	product_wg_ton	Product has been shipped in last 3 months. Weight is in tons

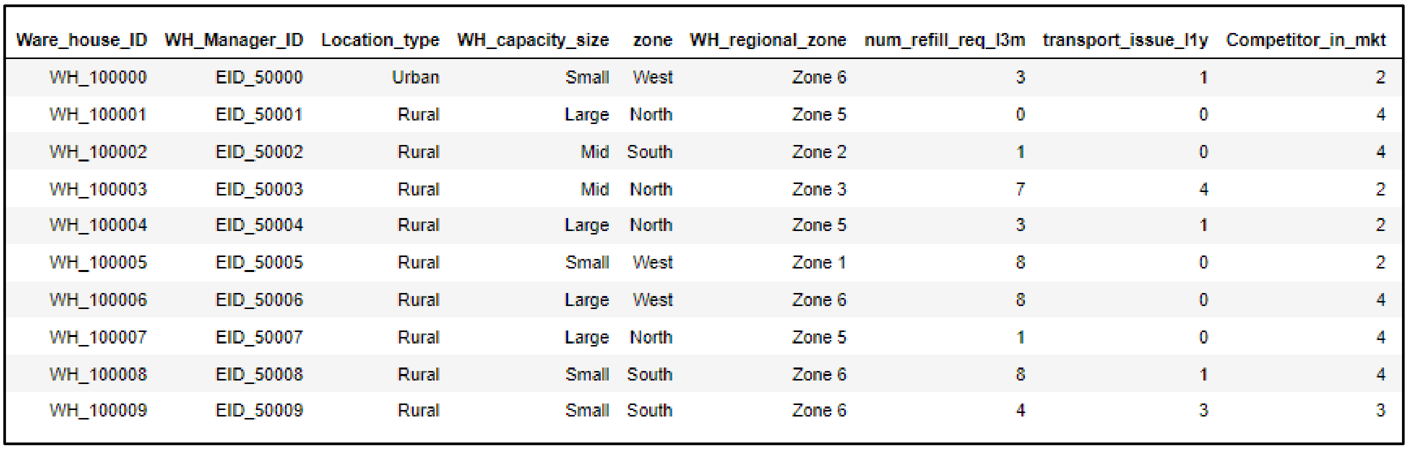
Table 1: Data dictionary for the dataset

There is total 24 variables in this dataset. It contains various measures related to company business.

Data report:

Visual inspection of data

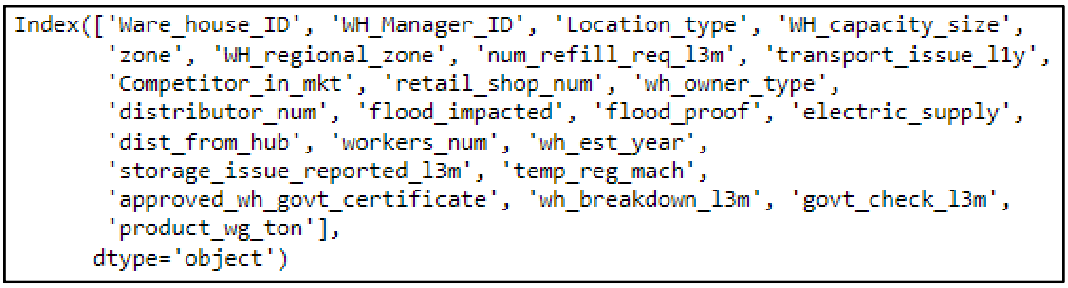
The dataset "Data.csv" is loaded using pandas and the dataset has 25,000 observations (rows) and 24 variables (columns). A quick glimpse of the data is shown below:



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
WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	2
WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	4
WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	4
WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	2
WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	2
WH_100005	EID_50005	Rural	Small	West	Zone 1	8	0	2
WH_100006	EID_50006	Rural	Large	West	Zone 6	8	0	4
WH_100007	EID_50007	Rural	Large	North	Zone 5	1	0	4
WH_100008	EID_50008	Rural	Small	South	Zone 6	8	1	4
WH_100009	EID_50009	Rural	Small	South	Zone 6	4	3	3

Figure 1: Dataset overview

Checking columns of dataset:



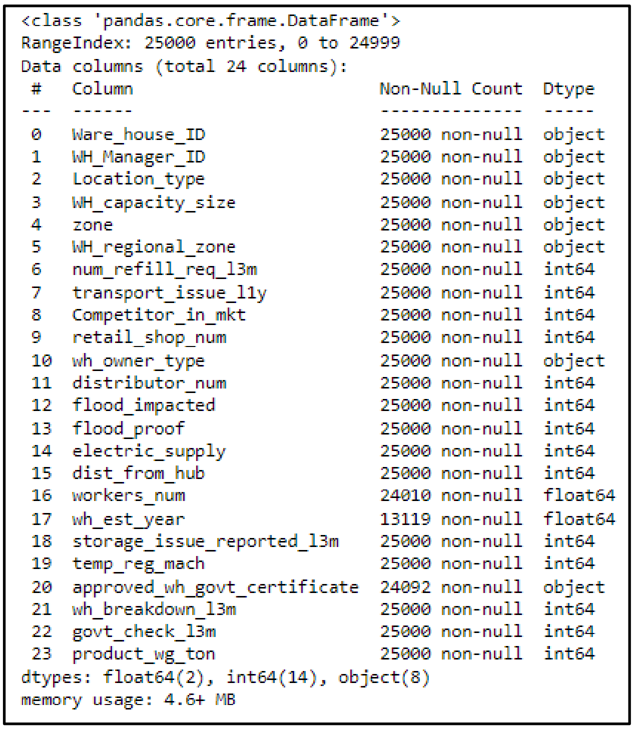
```
Index(['Ware_house_ID', 'WH_Manager_ID', 'Location_type', 'WH_capacity_size',
       'zone', 'WH_regional_zone', 'num_refill_req_l3m', 'transport_issue_l1y',
       'Competitor_in_mkt', 'retail_shop_num', 'wh_owner_type',
       'distributor_num', 'flood_impacted', 'flood_proof', 'electric_supply',
       'dist_from_hub', 'workers_num', 'wh_est_year',
       'storage_issue_reported_l3m', 'temp_reg_mach',
       'approved_wh_govt_certificate', 'wh_breakdown_l3m', 'govt_check_l3m',
       'product_wg_ton'],
      dtype='object')
```

Checking the shape of the data:

Number of rows: 25,000

Number of columns: 24

Description of variables are as below, to understand the data better:



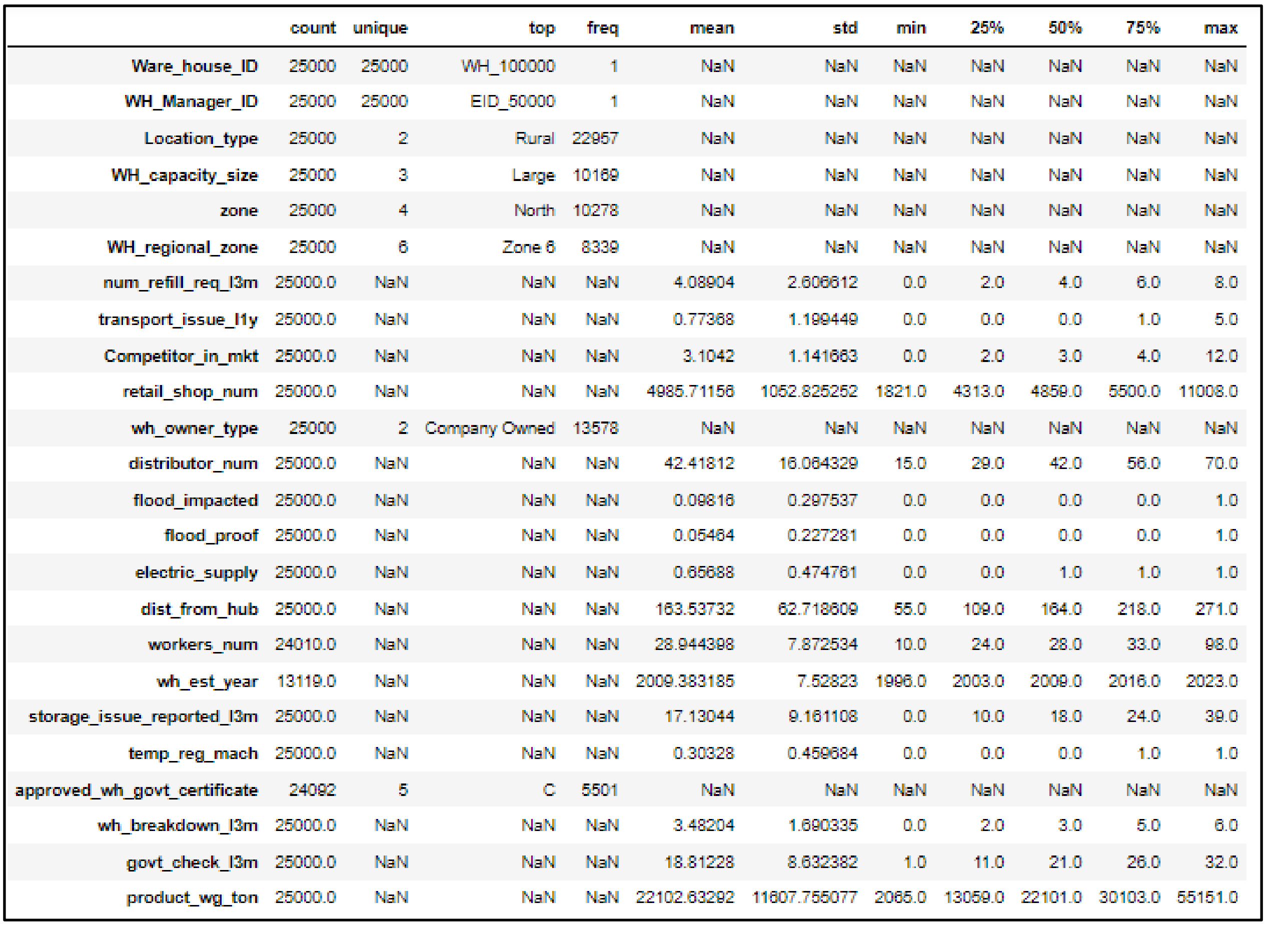
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Ware_house_ID                 25000 non-null  object 
 1   WH_Manager_ID                 25000 non-null  object 
 2   Location_type                 25000 non-null  object 
 3   WH_capacity_size              25000 non-null  object 
 4   zone                          25000 non-null  object 
 5   WH_regional_zone              25000 non-null  object 
 6   num_refill_req_l3m            25000 non-null  int64  
 7   transport_issue_l1y           25000 non-null  int64  
 8   Competitor_in_mkt             25000 non-null  int64  
 9   retail_shop_num               25000 non-null  int64  
 10  wh_owner_type                 25000 non-null  object 
 11  distributor_num               25000 non-null  int64  
 12  flood_impacted                25000 non-null  int64  
 13  flood_proof                   25000 non-null  int64  
 14  electric_supply               25000 non-null  int64  
 15  dist_from_hub                 25000 non-null  int64  
 16  workers_num                   24010 non-null  float64
 17  wh_est_year                   13119 non-null  float64
 18  storage_issue_reported_l3m    25000 non-null  int64  
 19  temp_reg_mach                 25000 non-null  int64  
 20  approved_wh_govt_certificate  24092 non-null  object 
 21  wh_breakdown_l3m              25000 non-null  int64  
 22  govt_check_l3m                25000 non-null  int64  
 23  product_wg_ton                25000 non-null  int64  
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Figure 2: Description of variables

Checking for Duplicate values in dataset:

Number of duplicate rows = 0

Statistical description of the dataset:

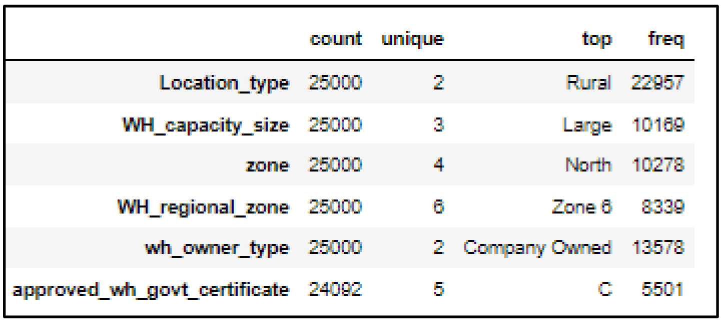


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ware_house_ID	25000	25000	WH_100000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_Manager_ID	25000	25000	EID_50000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Location_type	25000	2	Rural	22957	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_capacity_size	25000	3	Large	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zone	25000	4	North	10278	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_regional_zone	25000	6	Zone 6	8339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num_refill_req_l3m	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	NaN	NaN	NaN	4985.71156	1052.825252	1821.0	4313.0	4859.0	5500.0	11008.0
wh_owner_type	25000	2	Company Owned	13578	NaN	NaN	NaN	NaN	NaN	NaN	NaN
distributor_num	25000.0	NaN	NaN	NaN	42.41812	16.064329	15.0	29.0	42.0	56.0	70.0
flood_impacted	25000.0	NaN	NaN	NaN	0.09816	0.297537	0.0	0.0	0.0	0.0	1.0
flood_proof	25000.0	NaN	NaN	NaN	0.05464	0.227281	0.0	0.0	0.0	0.0	1.0
electric_supply	25000.0	NaN	NaN	NaN	0.65688	0.474761	0.0	0.0	1.0	1.0	1.0
dist_from_hub	25000.0	NaN	NaN	NaN	163.53732	62.718609	55.0	109.0	164.0	218.0	271.0
workers_num	24010.0	NaN	NaN	NaN	28.944398	7.872534	10.0	24.0	28.0	33.0	98.0
wh_est_year	13119.0	NaN	NaN	NaN	2009.383185	7.52823	1996.0	2003.0	2009.0	2016.0	2023.0
storage_issue_reported_l3m	25000.0	NaN	NaN	NaN	17.13044	9.161108	0.0	10.0	18.0	24.0	39.0
temp_reg_mach	25000.0	NaN	NaN	NaN	0.30328	0.459684	0.0	0.0	0.0	1.0	1.0
approved_wh_govt_certificate	24092	5	C	5501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wh_breakdown_l3m	25000.0	NaN	NaN	NaN	3.48204	1.690335	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	NaN	NaN	NaN	18.81228	8.632382	1.0	11.0	21.0	26.0	32.0
product_wg_ton	25000.0	NaN	NaN	NaN	22102.63292	11607.755077	2065.0	13059.0	22101.0	30103.0	55151.0

Figure 3: Statistical description of the dataset

The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile is mentioned in the above tables.

Summary statistics of the object variable:



	count	unique	top	freq
Location_type	25000	2	Rural	22957
WH_capacity_size	25000	3	Large	10169
zone	25000	4	North	10278
WH_regional_zone	25000	6	Zone 6	8339
wh_owner_type	25000	2	Company Owned	13578
approved_wh_govt_certificate	24092	5	C	5501

Observations:

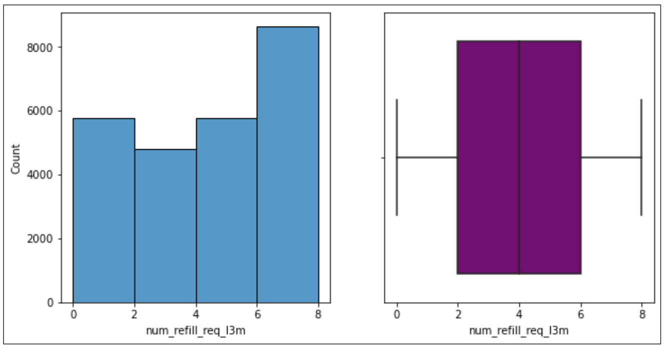
- The data set contains 25,000 rows 24 columns.
- 6 Categorical Variables, 4 Nominal Variables & 14 Continuous Variables.
- The product_wg_ton is the target variable, that is the objective of the study is to study and model the product weight in order to estimate the future demands.
- For our analysis, Ware_house_ID and WH_Manager_ID are dropped.
- There are no duplicate rows in the entire dataset.

2. EDA and Business Implication:

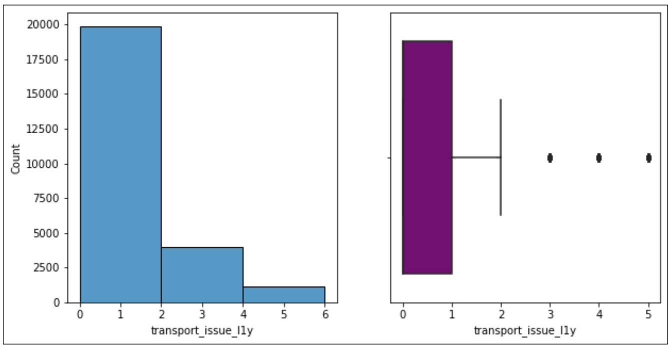
1. Univariate analysis:

Univariate analysis provides the distribution and spread for every continuous attribute, distribution of data in categories for categorical ones. Below charts provide the univariate analysis for the most critical variables:

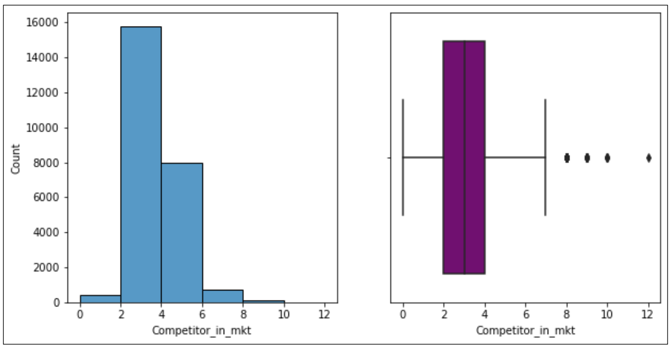
1. num_refill_req_l3m Variable:



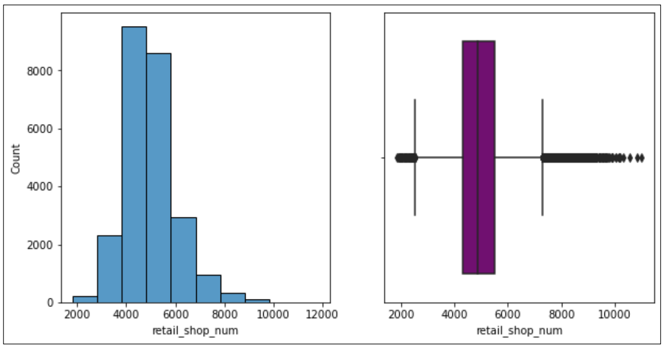
2. transport_issue_l1y Variable:



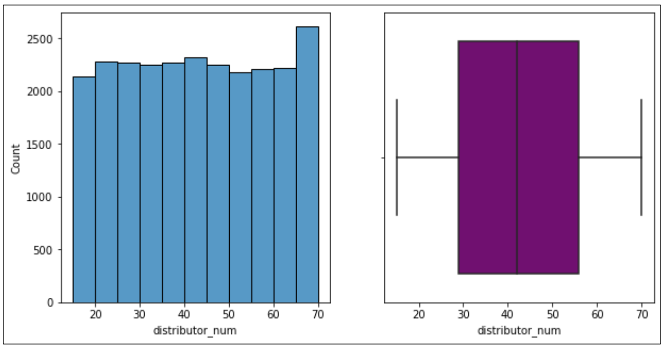
3. Competitor_in_mkt Variable:



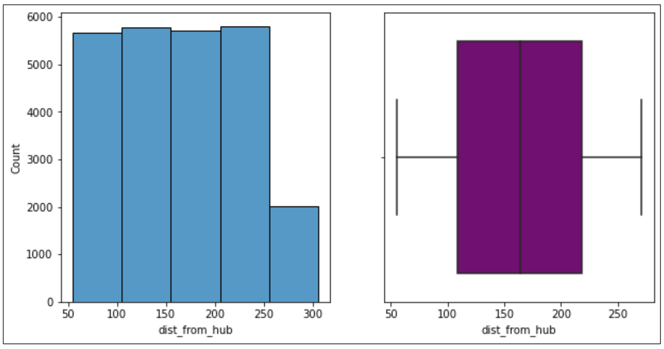
4. retail_shop_num Variable:



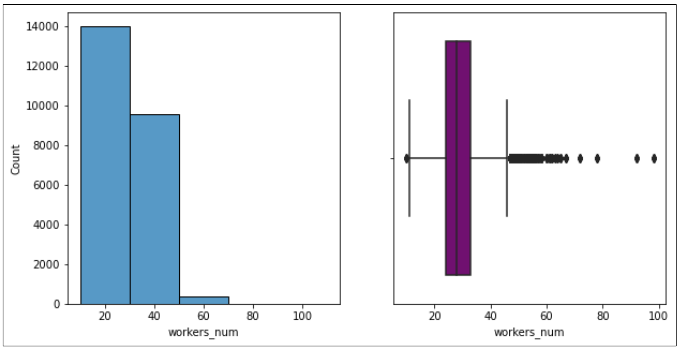
5. distributor_num Variable:



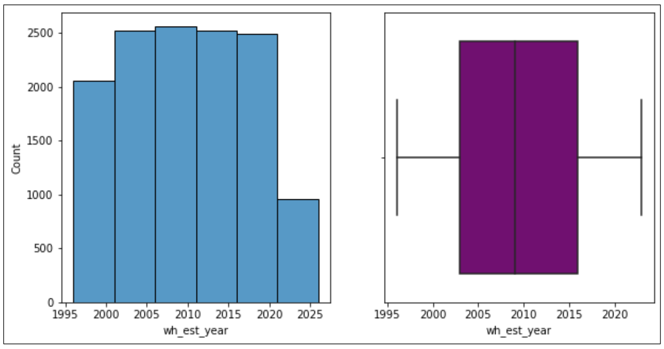
6. Dist_from_hub Variable:



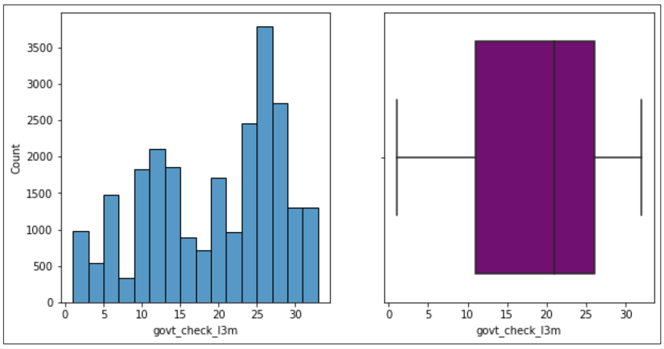
7. Workers_num Variable:



8. Wh_est_year Variable:



9. Govt_check_l3m Variable:



10. Product_wg_ton Variable:

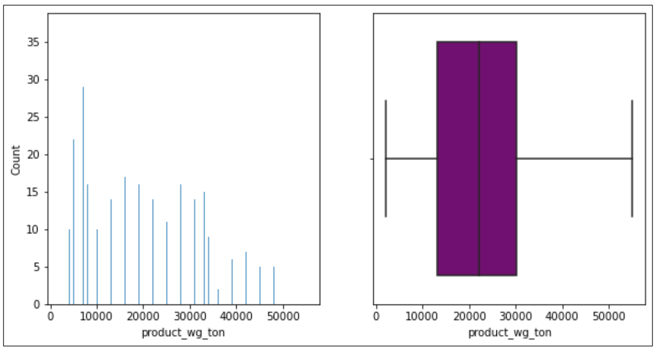


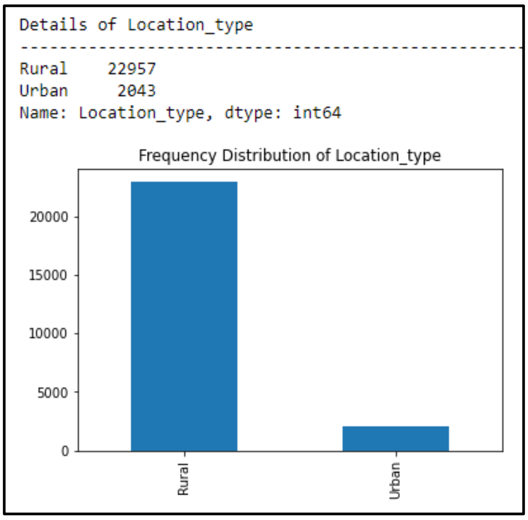
Figure 4: Univariate analysis

Observations:

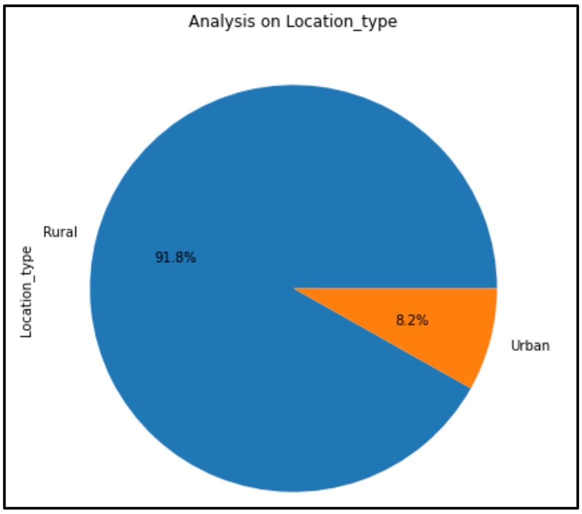
- Most of the variables are right skewed, left skewed variables are: wh_breakdown_13m, num_refill_req_13m, govt_check_13m, electric_supply.
- Retail shops are right skewed. Having 4000 to 5000 retail shops in normal.
- The number of distributors is observed to be fairly uniform across the respective ranges.
- The majority of warehouses have 20 to 40 workers.
- The number of storage issues reported is right skewed, with the majority having less than 20 issues across the duration of 3 months. The bin width is set to 10.
- It is observed that the majority of warehouses have reported more than 4 breakdowns in past 3 months. The data is left skewed.
- The number of government checks have been 25 to 30 for most warehouses.
- Distance from hub is observed to be fairly uniform across the respective ranges

Getting unique counts of Categorical Variables:

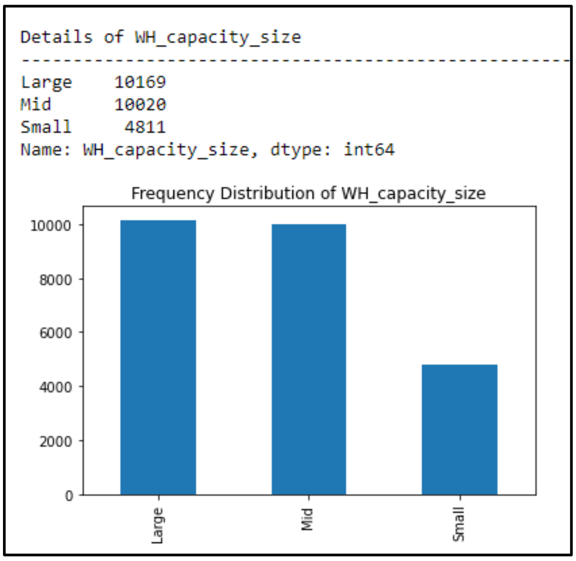
1. Location_type Variable:



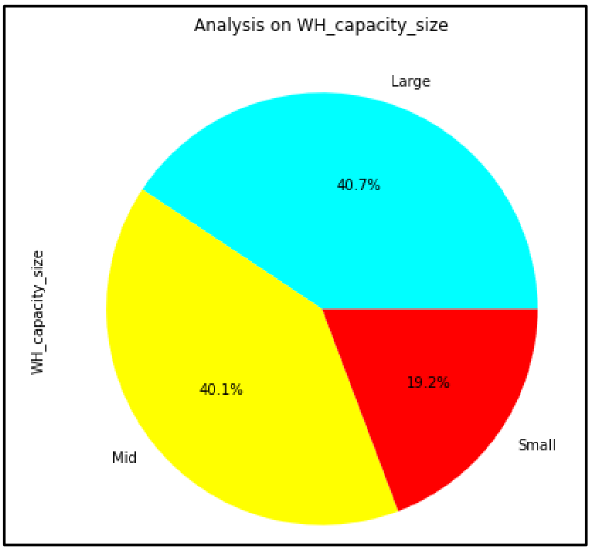
```
Details of Location_type
----------------------------------------------------
Rural     22957
Urban      2043
Name: Location_type, dtype: int64
```



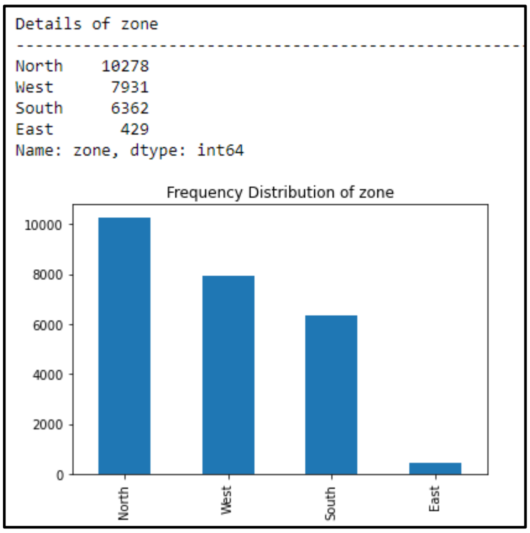
2. WH_capacity_size size Variable:



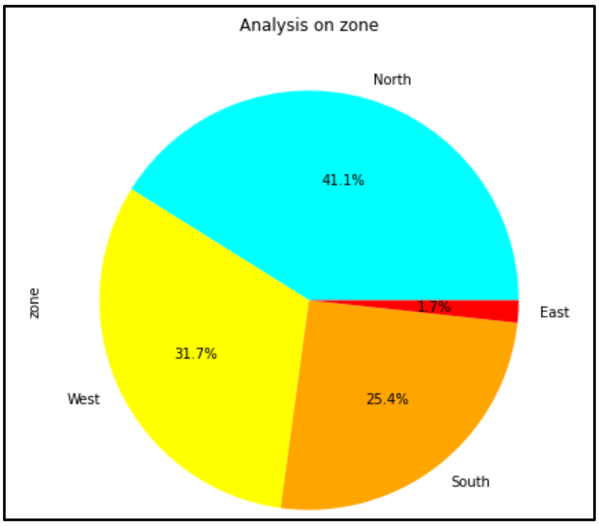
```
Details of WH_capacity_size
----------------------------------------------------
Large     10169
Mid       10020
Small      4811
Name: WH_capacity_size, dtype: int64
```



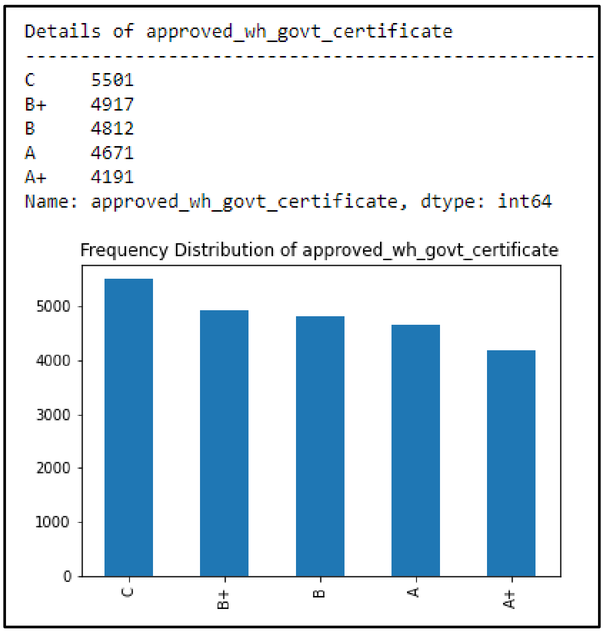
3. Zone Variable:



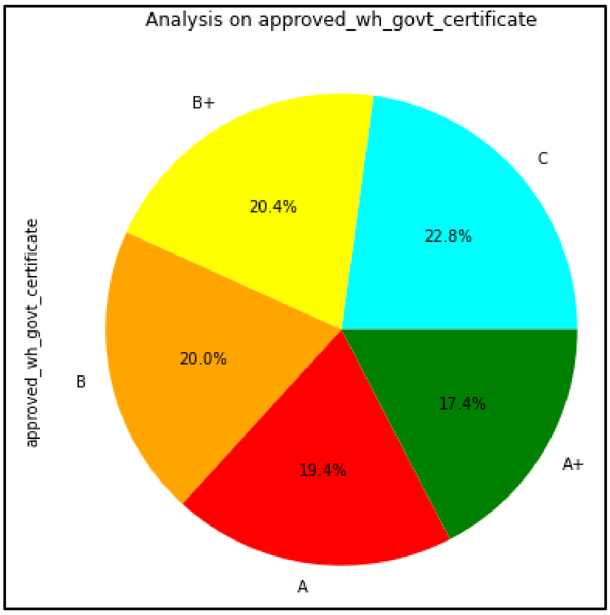
```
Details of zone
----------------------------------------------------
North    10278
West      7931
South     6362
East       429
Name: zone, dtype: int64
```



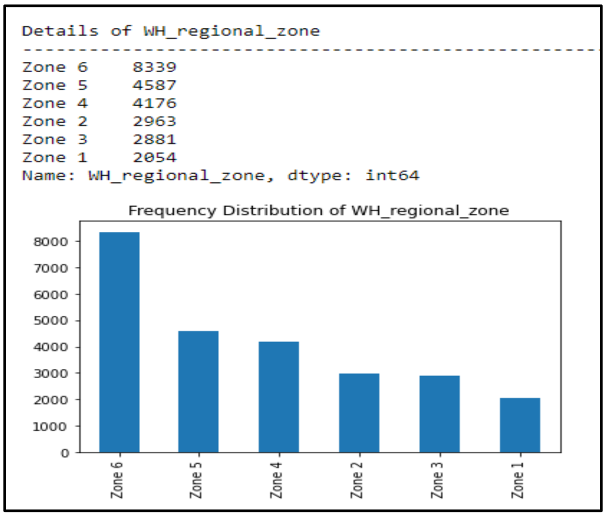
4. Approved_WH_Govt Variable:



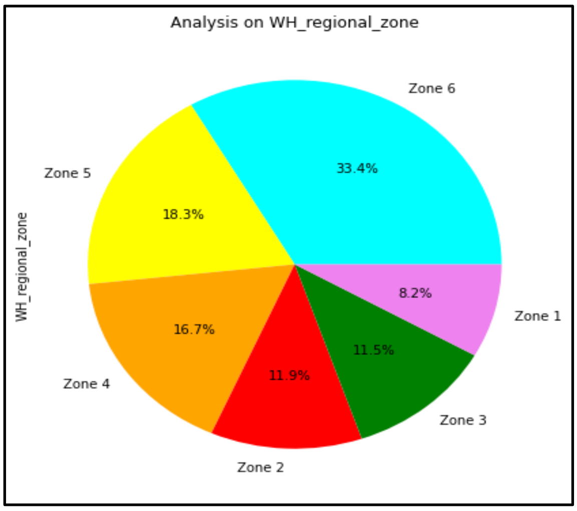
```
Details of approved_wh_govt_certificate
----------------------------------------------------------
C      5501
B+     4917
B      4812
A      4671
A+     4191
Name: approved_wh_govt_certificate, dtype: int64
```



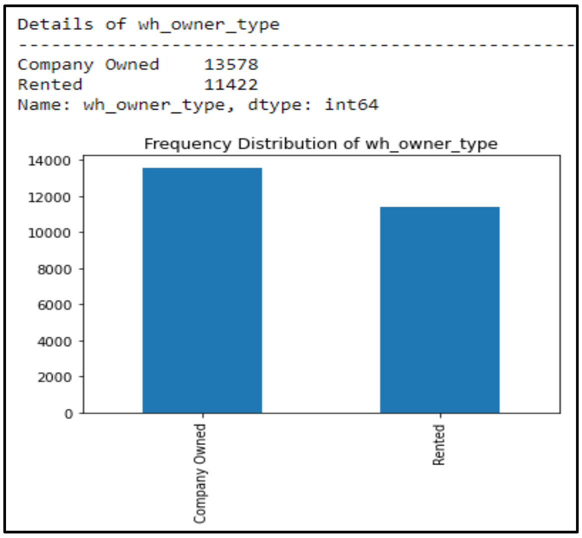
5. WH_regional_zone Variable:



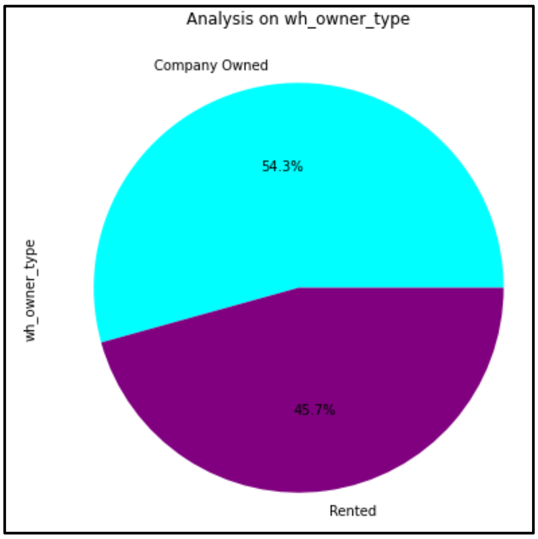
```
Details of WH_regional_zone
----------------------------------------------------------
Zone 6    8339
Zone 5    4587
Zone 4    4176
Zone 2    2963
Zone 3    2881
Zone 1    2054
Name: WH_regional_zone, dtype: int64
```



6. Owner type



```
Details of wh_owner_type
----------------------------------------------------------
Company Owned      13578
Rented             11422
Name: wh_owner_type, dtype: int64
```



Observations:

- Type of Location is Urban and Rural, with a higher count in rural type.
- Warehouse capacity are of 3 types: Small, Mid & large, it can be observed that small warehouses are less compared to that of Large and Mid.
- There are 2 Owner types: Rented & company owner. Company Owned warehouses are slightly more than the rented ones.
- The count of warehouses in the East zone are minimal compared to those in the North, West and South zones. majority of warehouses with electric supply have opted to go ahead without regulation of temperature facility. The warehouse also has a certificate (A+, A, B, B+ and C) based on the government standards.
- Regional zone numbered as Zone1 to Zone6, with Zone 1 has the lowest.
- It can be observed that the warehouses impacted by floods form a mere minority.

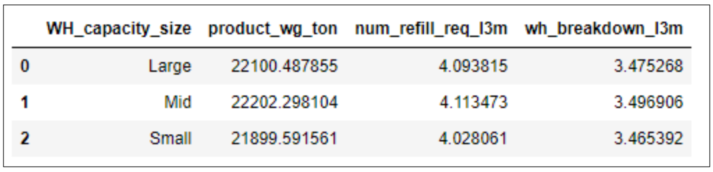
SKEWNESS VALUE:

Formula 1: **Skewness = 3 * (Mean – Median) / Standard Deviation**.

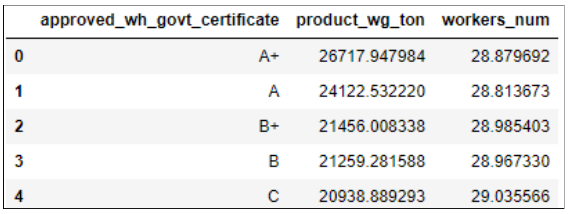
Variables	Skewness Value
flood_proof	3.92
flood_impacted	2.70
transport_issue_l1y	1.61
workers_num	1.06
Competitor_in_mkt	0.98
retail_shop_num	0.91
temp_reg_mach	0.86
product_wg_ton	0.33
storage_issue_reported_l3m	0.11
distributor_num	0.02
wh_est_year	0.01
dist_from_hub	-0.01
wh_breakdown_l3m	-0.07
num_refill_req_l3m	-0.08
govt_check_l3m	-0.36
electric_supply	-0.66

Table 2: Skewness values

2. Bivariate analysis:



	WH_capacity_size	product_wg_ton	num_refill_req_l3m	wh_breakdown_l3m
0	Large	22100.487855	4.093815	3.475268
1	Mid	22202.298104	4.113473	3.496906
2	Small	21899.591561	4.028061	3.465392



	approved_wh_govt_certificate	product_wg_ton	workers_num
0	A+	26717.947984	28.879692
1	A	24122.532220	28.813673
2	B+	21456.008338	28.985403
3	B	21259.281588	28.967330
4	C	20938.889293	29.035566

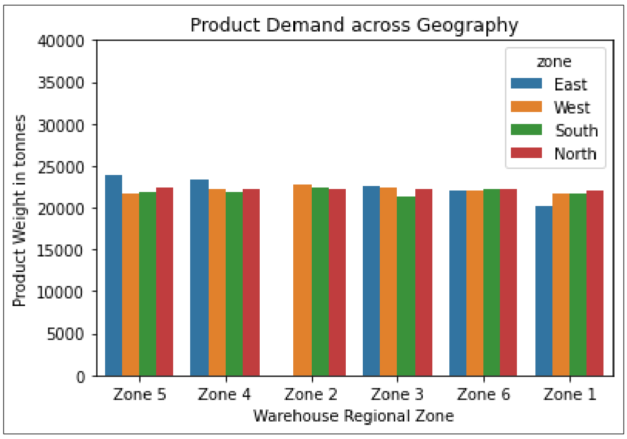


Figure 5: Bivariate analysis

It can be observed that East zone has no warehouses in Zone 2. In the North and the West, the Zone 6 has the highest number of warehouses. Further investigation shows that the East zone has fewer number of distributors and retail shops, but more competitors in the market, relative to the other zones.

In summary, this zone has fewer warehouses, retail outlets and distributors. It has much higher number of competitors yet has the same product demand as other zones. This might be a result of the popularity of the product in this region, encouraging the marketing department to pay greater attention to this region.

Correlation analysis:

Formula 2:

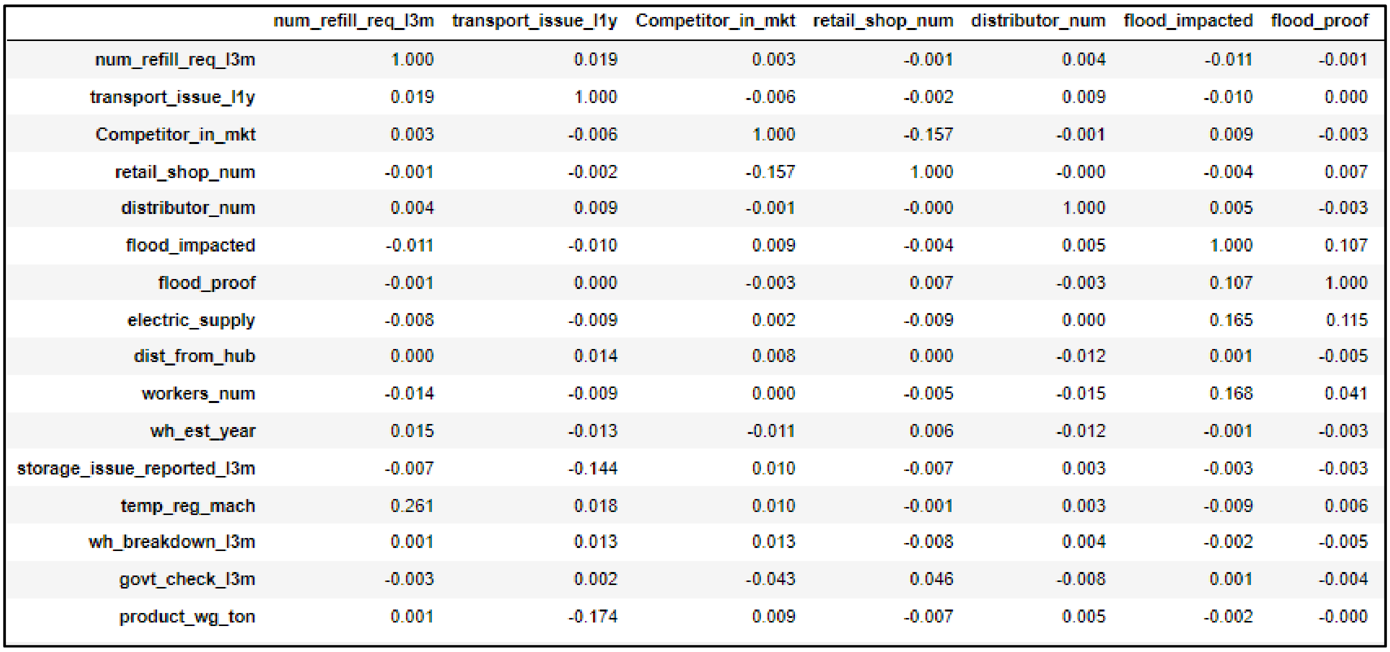
$$Correlation = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

Where,

$Cov(x, y) =$ Covariance of x and y

$\sigma_x =$ Standard deviation of x

$\sigma_y =$ Standard deviation of y



	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof
num_refill_req_l3m	1.000	0.019	0.003	-0.001	0.004	-0.011	-0.001
transport_issue_l1y	0.019	1.000	-0.006	-0.002	0.009	-0.010	0.000
Competitor_in_mkt	0.003	-0.006	1.000	-0.157	-0.001	0.009	-0.003
retail_shop_num	-0.001	-0.002	-0.157	1.000	-0.000	-0.004	0.007
distributor_num	0.004	0.009	-0.001	-0.000	1.000	0.005	-0.003
flood_impacted	-0.011	-0.010	0.009	-0.004	0.005	1.000	0.107
flood_proof	-0.001	0.000	-0.003	0.007	-0.003	0.107	1.000
electric_supply	-0.008	-0.009	0.002	-0.009	0.000	0.165	0.115
dist_from_hub	0.000	0.014	0.008	0.000	-0.012	0.001	-0.005
workers_num	-0.014	-0.009	0.000	-0.005	-0.015	0.168	0.041
wh_est_year	0.015	-0.013	-0.011	0.006	-0.012	-0.001	-0.003
storage_issue_reported_l3m	-0.007	-0.144	0.010	-0.007	0.003	-0.003	-0.003
temp_reg_mach	0.261	0.018	0.010	-0.001	0.003	-0.009	0.006
wh_breakdown_l3m	0.001	0.013	0.013	-0.008	0.004	-0.002	-0.005
govt_check_l3m	-0.003	0.002	-0.043	0.046	-0.008	0.001	-0.004
product_wg_ton	0.001	-0.174	0.009	-0.007	0.005	-0.002	-0.000

Figure 6: Correlation analysis

Correlation Heatmap:

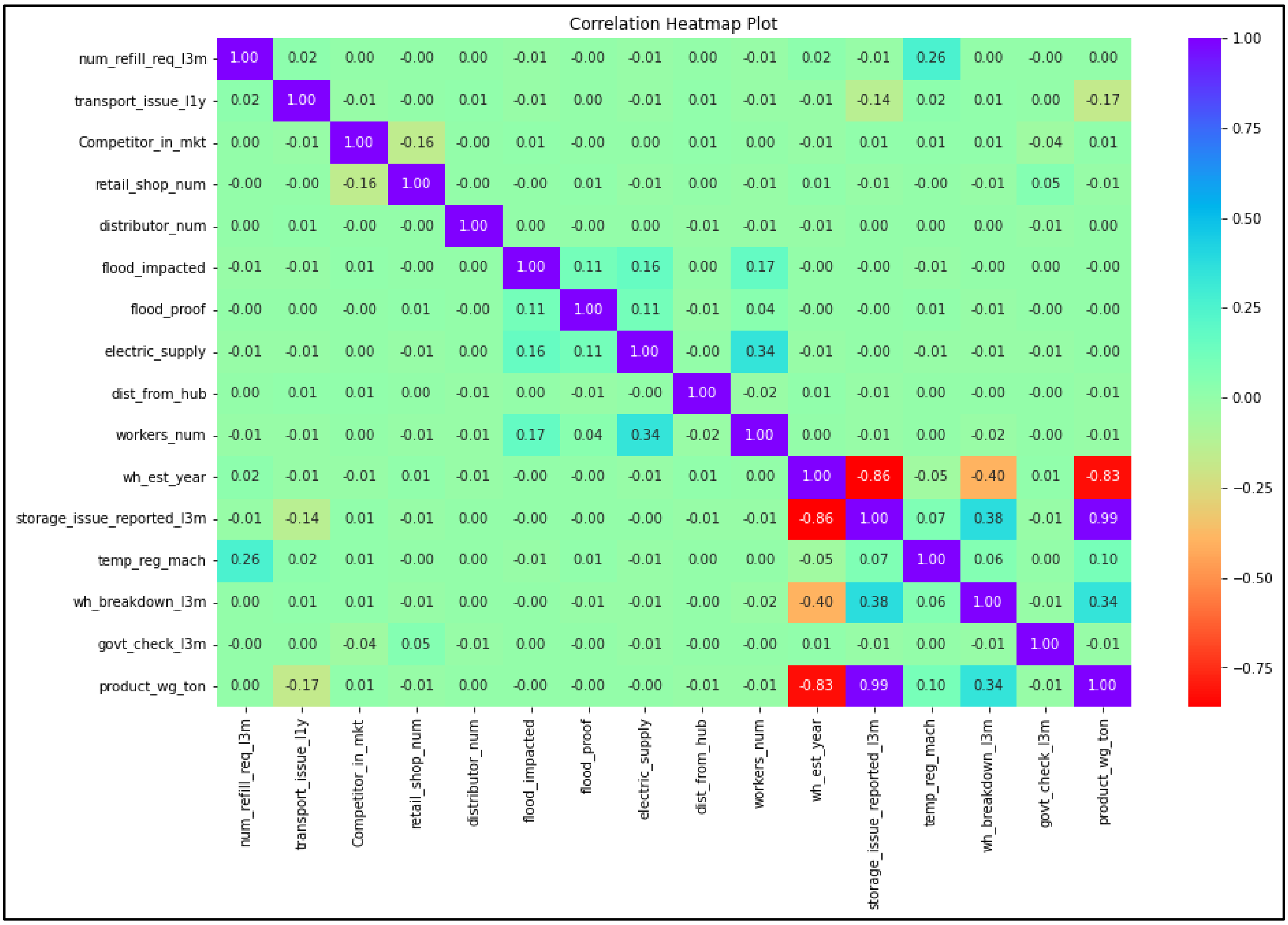


Figure 7: Correlation heatmap

❖ It is observed that the product demand has a strong positive correlation with the storage issues reported.

- The Pearson correlation coefficients calculated for all pairs of continuous variables obtained as a matrix, have been presented as a heatmap. The darker the colour, the lower is the magnitude of the correlation.

- As can be observed, “storage issues reported in 3 months” column has a high correlation with product weight in tons.

Pair plot:

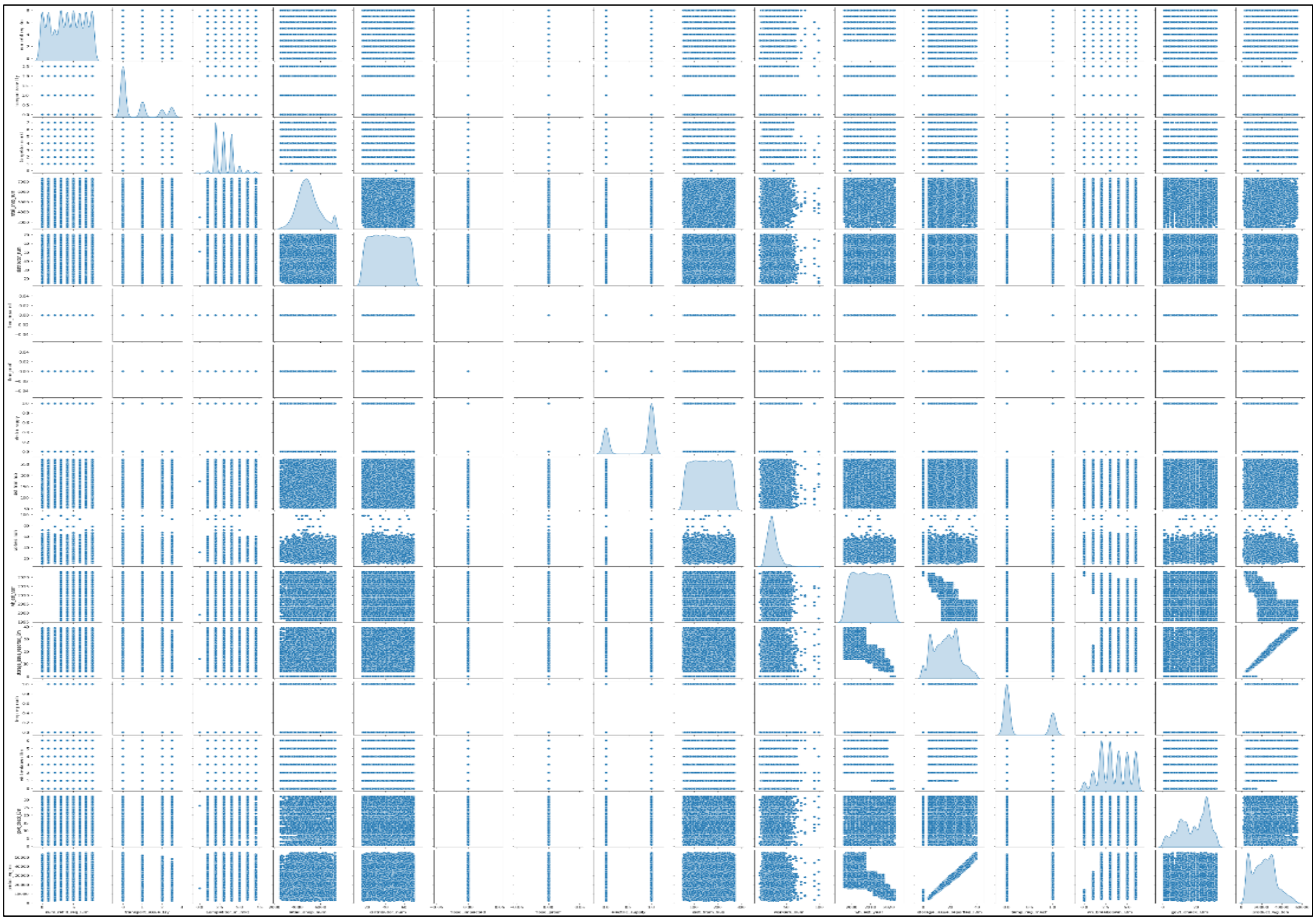


Figure 8: Pair plot analysis

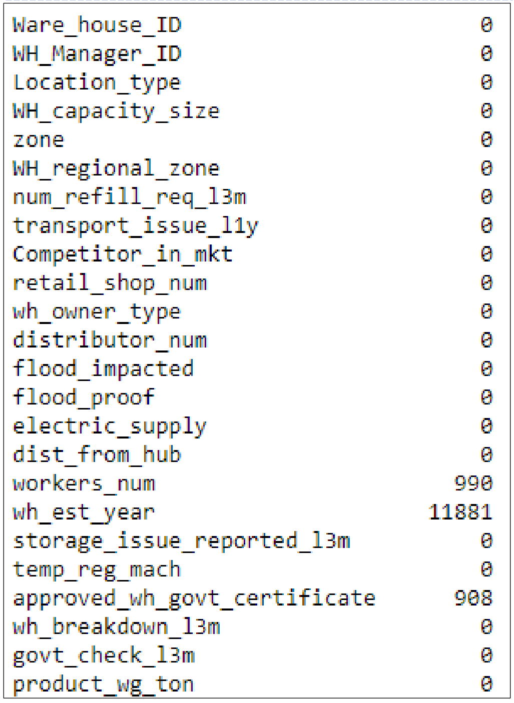
3. Data Cleaning and Pre-processing

Removal of unwanted variables:

There are two ID variables: Warehouse ID and Warehouse Manager ID. These columns have not been included in this analysis.

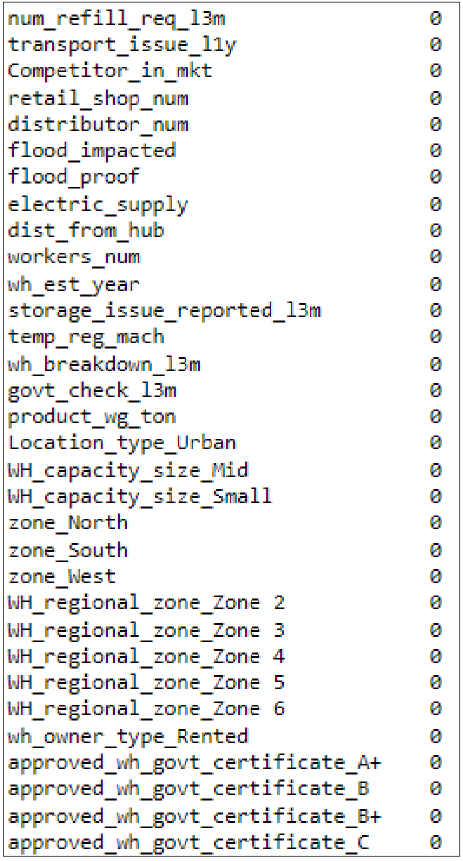
Missing Value treatment:

Checking the missing value among all variables



```
Ware_house_ID                    0
WH_Manager_ID                    0
Location_type                    0
WH_capacity_size                 0
zone                             0
WH_regional_zone                 0
num_refill_req_l3m               0
transport_issue_l1y              0
Competitor_in_mkt                0
retail_shop_num                  0
wh_owner_type                    0
distributor_num                  0
flood_impacted                   0
flood_proof                      0
electric_supply                  0
dist_from_hub                    0
workers_num                    990
wh_est_year                  11881
storage_issue_reported_l3m       0
temp_reg_mach                    0
approved_wh_govt_certificate   908
wh_breakdown_l3m                 0
govt_check_l3m                   0
product_wg_ton                   0
```

Applied the KNN imputer, to remove the missing values in the dataset



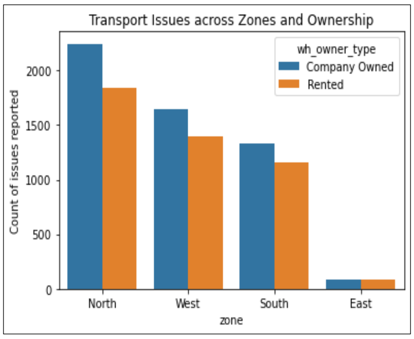
```
num_refill_req_l3m                0
transport_issue_l1y               0
Competitor_in_mkt                 0
retail_shop_num                   0
distributor_num                   0
flood_impacted                    0
flood_proof                       0
electric_supply                   0
dist_from_hub                     0
workers_num                       0
wh_est_year                       0
storage_issue_reported_l3m        0
temp_reg_mach                     0
wh_breakdown_l3m                  0
govt_check_l3m                    0
product_wg_ton                    0
Location_type_Urban               0
WH_capacity_size_Mid              0
WH_capacity_size_Small            0
zone_North                        0
zone_South                        0
zone_West                         0
WH_regional_zone_Zone 2           0
WH_regional_zone_Zone 3           0
WH_regional_zone_Zone 4           0
WH_regional_zone_Zone 5           0
WH_regional_zone_Zone 6           0
wh_owner_type_Rented              0
approved_wh_govt_certificate_A+   0
approved_wh_govt_certificate_B    0
approved_wh_govt_certificate_B+   0
approved_wh_govt_certificate_C    0
```

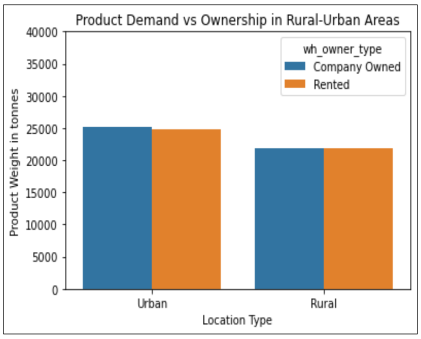
Figure 9: Missing value treatment

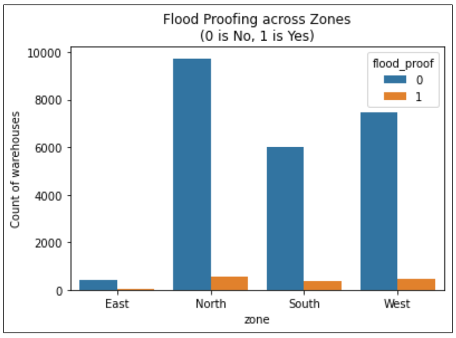
Observations:

- 'workers_num' variable has 990 missing values.
- 'wh_est_year' variable has 11881 missing values.
- 'approved_wh_govt_certificate' variable has 908 missing values.
- In total, 2.3% of the data has null values.
- Post KNN imputation method, all missing values have been removed.
- The number of nearest neighbours to be considered is kept at 1. However, as they are categorical in nature, these are converted to dummy variables eventually. The output after the KNN imputation is shown above (right side)

Additional observations:





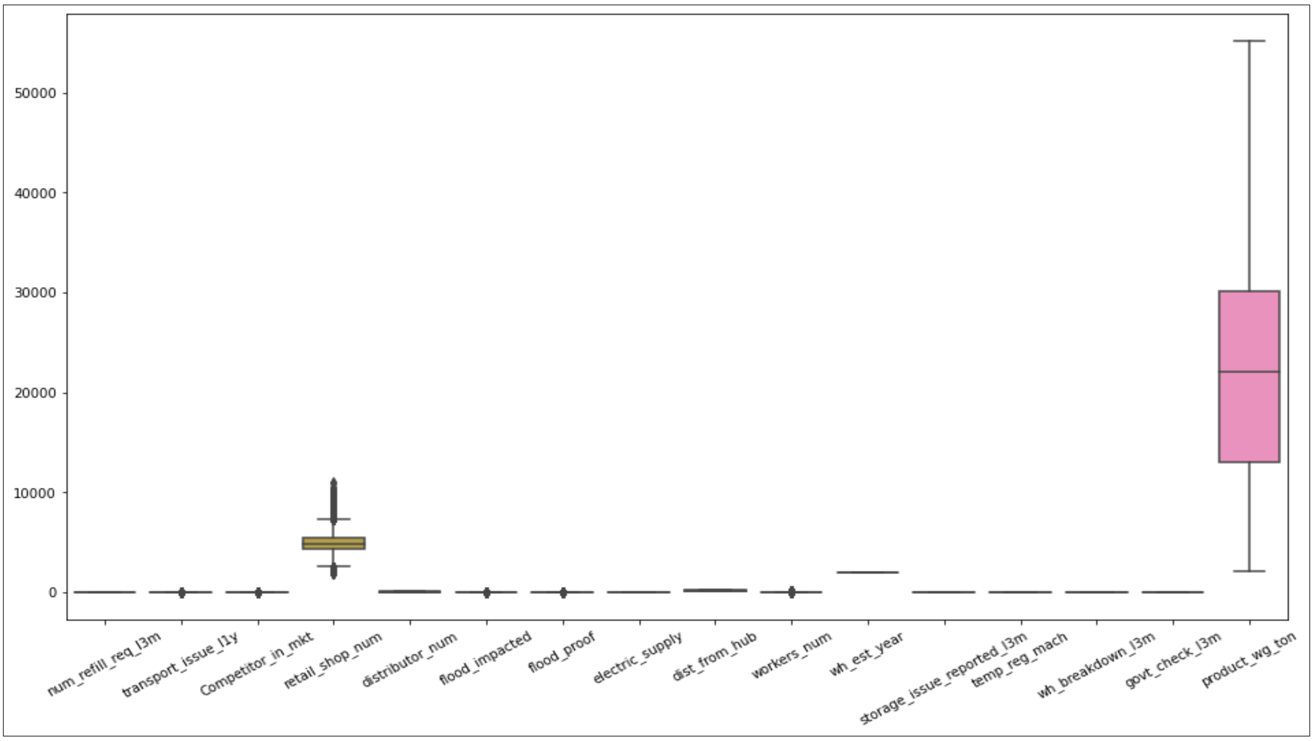


- The ownership of the warehouse seems to have an impact on the average product weight shipped, which also varies based on the location of the warehouse. The urban company-owned warehouses order more than the rented ones in the same location. In rural areas, this difference is negligible.
- A greater number of transport issues have been reported by company-owned warehouses, in comparison to the rented ones. This trend is consistent in all zones, except the East, were there is no difference.
- Flood proofing though a desirable facility at any storage facility, it can be observed that the warehouses impacted by floods form a mere minority. This justifies why warehouses have not opted for flood proofing.

Outlier treatment:

The descriptive analysis of the data has been followed by outlier detection and removal. This is limited to continuous variables. Below are the variables having outliers.

Before treating outliers:



The IQR (Interquartile Range) method has been used to detect and remove outliers. IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

For the treatment of such outliers, the values lesser than lower bound are replaced with the value of lower bound, and the values greater than upper bound are replaced by the value of upper bound. Thus, the range of the values shrinks.

After treating outliers:

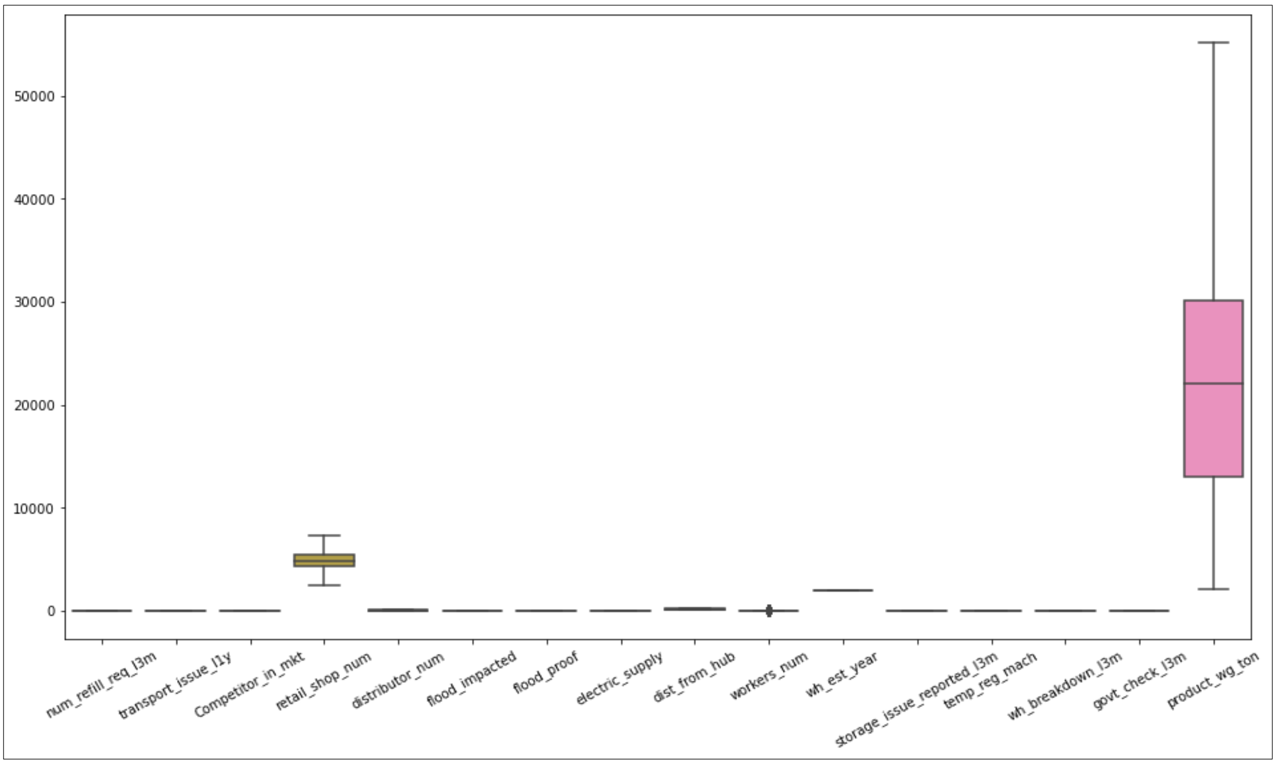


Figure 10: Outlier analysis

- Competitor_in_mkt, retail_shop_num, transport_issue_l1y, workers_num, the outlier treatment is done in the further analysis.
- Flood_impacted & flood_proof outliers are not treated as they are nominal columns.
- It can be observed that in the “product_wg_ton” variable, outliers could not be removed. However, there is no impact in the overall analysis.

Addition of new variables - Age group:

- Created a new variable, “Age group” by using imputation and lambda function
- This is used to create bins -> [0,5,10,15,20,25]
- Below is the distribution across the bins of the age group.

0-5	5,155
5-10	4,720
10-15	5,789
15-20	4,411
20-25	4,263

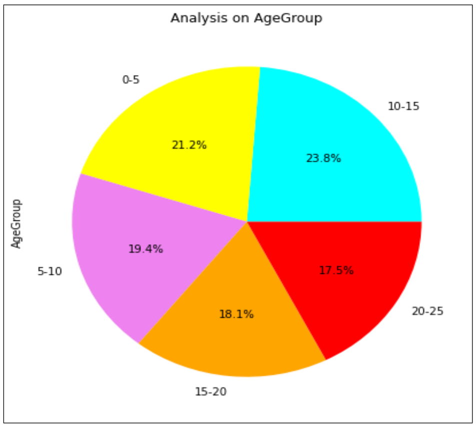


Figure 11: Additional of new variables

Business insights from EDA:

Data Imbalance and its Treatment:

Data imbalance is applicable to the classification problem. Since this is a Regression problem, we no need do any data Imbalance treatment here. Clustering is done for unsupervised data, when no historical data is given. Here in our study, we are given with the historical data and it is a supervised problem, hence clustering is not applicable.

Business insights:

A. Warehouse location:

- ❖ More warehouses are situated in the rural regions, than urban.
- ❖ There are unusually low number of warehouses in the East region. This region has more competition and more scope of growth. Thus, the marketing team can focus on this region. The legal team needs to figure out why this place gets more government checks than other regions.

B. Warehouse features:

- ❖ One third of the warehouses have no electricity and hence this needs close monitoring during product storage to avoid any perishability.
- ❖ A minority have opted for flood proofing and temperature regulation. The ones with a temperature regulator, report more breakdowns and hence the quality of these machines should be improved.
- ❖ The warehouses with A+ certifications have on average, greater demand for products.

C. Warehouse ownership:

- ❖ The ownership of the warehouse seems to have an impact on the average product weight shipped, which also varies based on the warehouse location. The urban company owned warehouses order more than the rented ones in the same location.
- ❖ A greater number of transportation issues have been reported by company-owned warehouses, in comparison to the rented ones. This trend is consistent across all zones, except the East, were there is no difference. This could be an additional area of improvement.

4. Model building:

Build various models:

Feature Selection:

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. Feature selection can be done in multiple ways but there are broadly 3 categories:

1. Filter Method
2. Wrapper Method
3. Embedded Method.

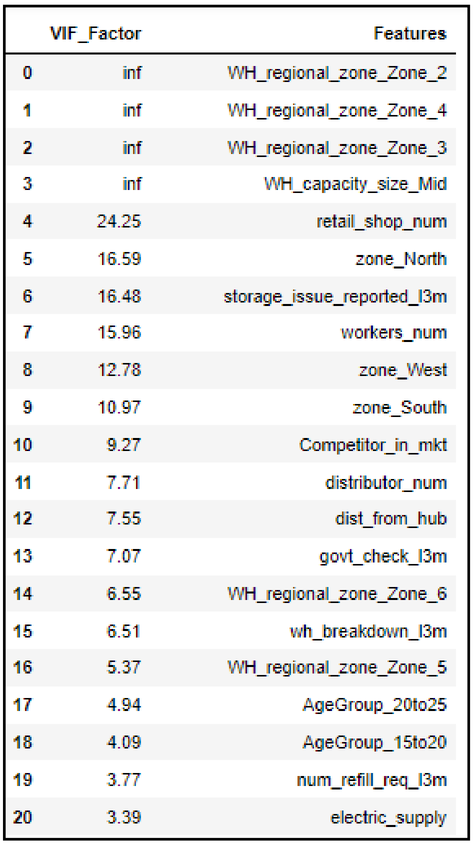
1. Filter Method

In this method you filter and take only the subset of the relevant features. The model is built after selecting the features. The filtering here is done using correlation matrix and it is most commonly done using Pearson correlation and VIF.

Variance Inflation Factor (VIF):

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. Collinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset.

Formula 3: $VIF = \frac{1}{1-r^2}$



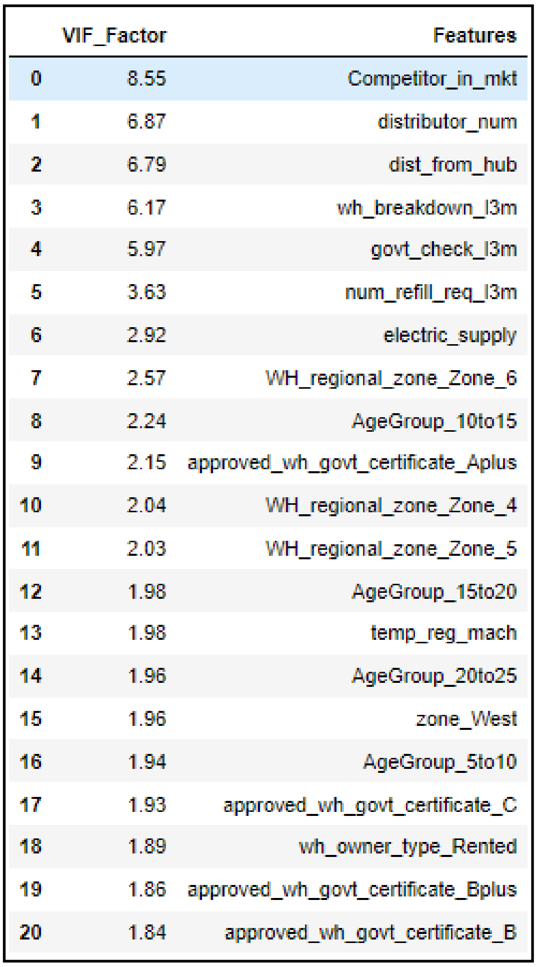
	VIF_Factor	Features
0	inf	WH_regional_zone_Zone_2
1	inf	WH_regional_zone_Zone_4
2	inf	WH_regional_zone_Zone_3
3	inf	WH_capacity_size_Mid
4	24.25	retail_shop_num
5	16.59	zone_North
6	16.48	storage_issue_reported_l3m
7	15.96	workers_num
8	12.78	zone_West
9	10.97	zone_South
10	9.27	Competitor_in_mkt
11	7.71	distributor_num
12	7.55	dist_from_hub
13	7.07	govt_check_l3m
14	6.55	WH_regional_zone_Zone_6
15	6.51	wh_breakdown_l3m
16	5.37	WH_regional_zone_Zone_5
17	4.94	AgeGroup_20to25
18	4.09	AgeGroup_15to20
19	3.77	num_refill_req_l3m
20	3.39	electric_supply

Figure 12: Variance Inflation Factor

We set the threshold to 10, as we wish to remove the variable for which the remaining variables explain more than 90% of the variation. Now we will remove the variables at the top "WH_regional_zone_Zone_2, retail_shop_num, workers_num, storage_issue_reported_l3m, zone North" as its still showing VIF value greater than 10.

Here we perform the VIF and will remove the variables one by one which are highly correlated and proceeding with the other variable for modelling.

After few iterations the VIF, below are the variable selected to build the model.

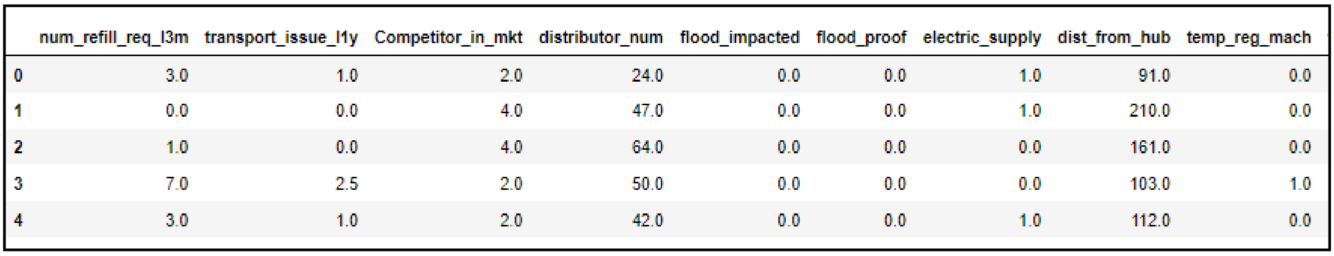


	VIF_Factor	Features
0	8.55	Competitor_in_mkt
1	6.87	distributor_num
2	6.79	dist_from_hub
3	6.17	wh_breakdown_l3m
4	5.97	govt_check_l3m
5	3.63	num_refill_req_l3m
6	2.92	electric_supply
7	2.57	WH_regional_zone_Zone_6
8	2.24	AgeGroup_10to15
9	2.15	approved_wh_govt_certificate_Aplus
10	2.04	WH_regional_zone_Zone_4
11	2.03	WH_regional_zone_Zone_5
12	1.98	AgeGroup_15to20
13	1.98	temp_reg_mach
14	1.96	AgeGroup_20to25
15	1.96	zone_West
16	1.94	AgeGroup_5to10
17	1.93	approved_wh_govt_certificate_C
18	1.89	wh_owner_type_Rented
19	1.86	approved_wh_govt_certificate_Bplus
20	1.84	approved_wh_govt_certificate_B

Figure 13: VIF for selected variables

Now all the variables above have VIF values less than 10, will continue with these variables and so the model building.

Below is the head of the dataset after dropping the columns with VIF>10:



	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_hub	temp_reg_mach
0	3.0	1.0	2.0	24.0	0.0	0.0	1.0	91.0	0.0
1	0.0	0.0	4.0	47.0	0.0	0.0	1.0	210.0	0.0
2	1.0	0.0	4.0	64.0	0.0	0.0	0.0	161.0	0.0
3	7.0	2.5	2.0	50.0	0.0	0.0	0.0	103.0	1.0
4	3.0	1.0	2.0	42.0	0.0	0.0	1.0	112.0	0.0

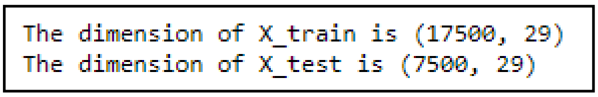
Figure 14: VIF Head data for selected variables

Train -Test Split

After selecting the variable for model building, we have performed the train test split.

- X= Copy all the predictor variables & y= target into the y data frame.
- Splitting the X and y into training and test set in 70:30 ratio with random_state=1.

Dimension of train and test dataset:



```
The dimension of X_train is (17500, 29)
The dimension of X_test is (7500, 29)
```

Scaling:

Data standardization is the process where using which we bring all the data under the same scale. Here, we are building a model, to predict optimum weight of the product to be shipped each time tothe warehouse. In this case we are expected to build model using Linear Regression, LDA, Ridge, Lasso, ANN etc. So, we are scaling the data (x_train_scaled, x_test_scaled) and will use this scaled data to perform the models where scaling is necessary.

Models

Since this a supervised regression problem will be performing some of the regression models below.Two metrics that statisticians often use to quantify how well a model fits a dataset are the root mean squared error (RMSE) and the R-squared (R2).

Predictive model against the test set using various appropriate performance metrics:

Linear Regression:

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable. It is mostly used for finding out the relationship between variables and forecasting.

Formula 4: **Linear Regression:**

$$Y = m_1X_1 + m_2X_2 + \cdots + m_nX_n + C + e$$

Y = Dependent / target / predicted variable
X_i= Independent/ predictor variable
m_i= coefficients for the i^{th} independent / predictor variable.
C = constant / intercept / bias
e = residual error / unexplained variance / difference between actual and prediction.

Test Results:

	RMSE	Accuracy Score
Train	6307.43	0.71
Test	6110.81	0.72

ANN Regressor:

Artificial Neural Networks have the ability to learn the complex relationship between the features and target due to the presence of activation function in each layer.

Test Results:

	RMSE	Accuracy Score
Train	6270.43	0.71
Test	6075.72	0.72

After running the base models, below are the RMSE and R2 score values:

	Train (RMSE)	Test (RMSE)	Training (R2)	Test (R2)
Ridge Regression	6307.43	6110.81	0.71	0.72
Lasso Regression	6307.43	6110.81	0.71	0.72
Linear Regression	6307.43	6110.81	0.71	0.72
Decision Tree Regressor	0.00	8231.95	1.00	0.49
Random Forest Regressor	2267.59	5918.08	0.96	0.74
ANN Regressor	6270.43	6075.72	0.71	0.72

Table 3: Base Model Results

Interpretation of the model:

- We can observe that the results are almost similar for Linear, Lasso and Ridge regression. Hence the features are selected using VIF method, Lasso and Ridge are performing same as linear regression.
- Decision tree and Random Forest's nonlinear nature gives better results than linear regression.

  Decision tree's accuracy shows that it is overfitting, so does random forest's results show.
- Linear regression and other methods can understand only linear relationships, to understand non-linear relationships ANN works better. Looking at the result ANN performs better than Linear and regularization methods. Real life data is supposed to have complex non-linear relationships, that's why ANN is giving better results than linear model.
- From the models it can be inferred that warehouse established year, number of refills, warehouse breakdown, distribution from hub etc. are few of the importance features effecting the optimum weight shipment.
- We can use grid search to tackle this problem Later, we will try to tune the models and will see whether the model performance improves.

Model Tuning and business implication

Ensemble modelling:

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. Tried using few of the ensemble modelsbelow to see whether the model performs better than base models.

Bagging regressor:

It is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Test Results:

	RMSE	Accuracy Score
Train	2684.79	0.95
Test	6150.14	0.71

AdaBoost regressor:

It is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. It is best used with weak learners.

Test Results:

	RMSE	Accuracy Score
Train	6942.13	0.64
Test	6849.01	0.65

Gradient Boost regressor:

Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learnersor weak predictive models.

Test Results:

	RMSE	Accuracy Score
Train	5967.89	0.74
Test	5819.52	0.74

Extreme Gradient Boosting:

XGBoost is a powerful approach for building supervised regression models. It is an efficient implementation of gradient boosting that can be used for regression predictive modelling.

Test Results:

	RMSE	Accuracy Score
Train	4380.68	0.86
Test	5991.81	0.73

Ensemble Model results as below:

	Train RMSE	Test RMSE	Training Score	Test Score
AdaBoost Regressor	6942.13	6849.01	0.64	0.65
Gradient Boosting Regressor	5967.89	5819.52	0.74	0.74
Bagging Regressor	2684.79	6150.14	0.95	0.71
XGB Regressor	4380.68	5991.81	0.86	0.73

Table 4: Ensemble Model results

Insights:

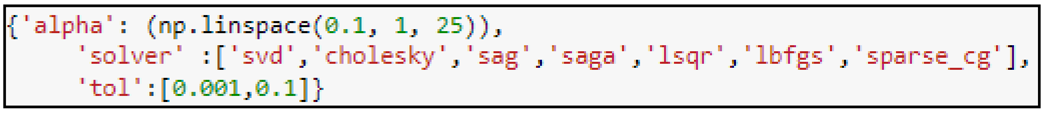
- Both Bagging & XGB regressor models are not performing well.
- Within all the ensemble models shown above Gradient Boosting regressor is performing better.From
- the models we could see that warehouse established year, transport issue, warehouse breakdown is few of the importance features effecting the optimum weight shipment.
- Later, we will try to tune the models and will see whether the model performance improves.

Model tuning measures:

Model Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. This is accomplished by selecting appropriate "hyperparameters", these parameters are set manually. The three most commonly used approaches are Grid Search, Random Search & K-fold. Here GridSearchCV Method is used for the model tuning.

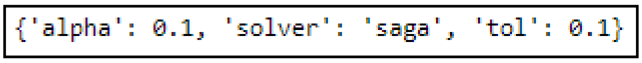
GridSearchCV on Ridge Regression

The given tuning parameters are below:



```
{'alpha': (np.linspace(0.1, 1, 25)),
     'solver' :['svd','cholesky','sag','saga','lsqr','lbfgs','sparse_cg'],
     'tol':[0.001,0.1]}
```

The best parameters after fitting the model are:



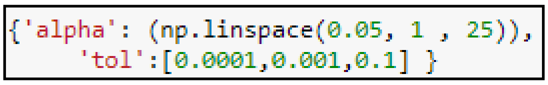
```
{'alpha': 0.1, 'solver': 'saga', 'tol': 0.1}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	6560.29	0.68
Test	6398.49	0.69

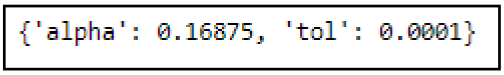
Lasso Regressor with GridSearchCV:

The given tuning parameters are below:



```
{'alpha': (np.linspace(0.05, 1 , 25)),
    'tol':[0.0001,0.001,0.1] }
```

The best parameters after fitting the model are:



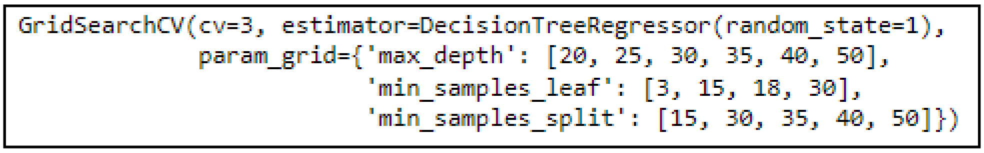
```
{'alpha': 0.16875, 'tol': 0.0001}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	6308.69	0.71
Test	6110.29	0.72

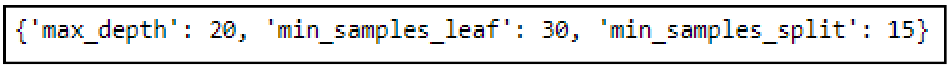
GridSearchCV on Decision Tree

Building a Decision Tree Regressor using a grid search cross validation to get best parameter orestimators for a given dataset. The given tuning parameters are below:



```
GridSearchCV(cv=3, estimator=DecisionTreeRegressor(random_state=1),
             param_grid={'max_depth': [20, 25, 30, 35, 40, 50],
                         'min_samples_leaf': [3, 15, 18, 30],
                         'min_samples_split': [15, 30, 35, 40, 50]})
```

The best parameters after fitting the model are:



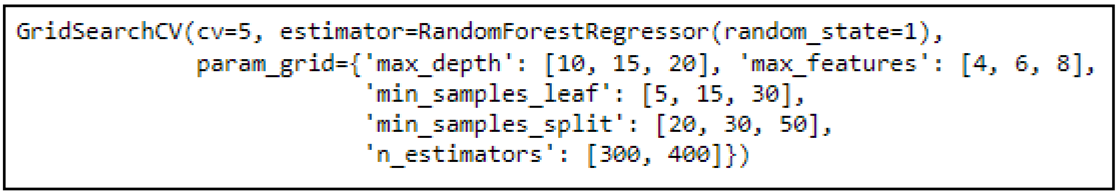
```
{'max_depth': 20, 'min_samples_leaf': 30, 'min_samples_split': 15}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	5527.60	0.77
Test	5933.77	0.73

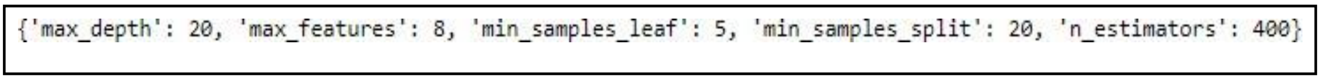
GridSearchCV on Random Forest Regressor:

The given tuning parameters are below:



```
GridSearchCV(cv=5, estimator=RandomForestRegressor(random_state=1),
             param_grid={'max_depth': [10, 15, 20], 'max_features': [4, 6, 8],
                         'min_samples_leaf': [5, 15, 30],
                         'min_samples_split': [20, 30, 50],
                         'n_estimators': [300, 400]})
```

The best parameters after fitting the model are:



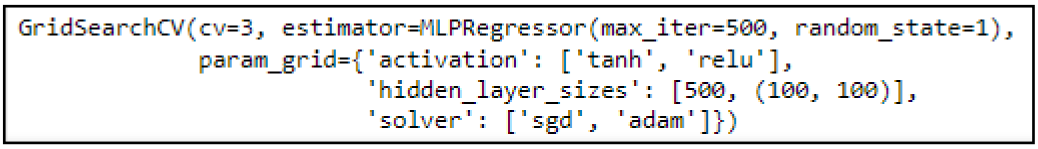
```
{'max_depth': 20, 'max_features': 8, 'min_samples_leaf': 5, 'min_samples_split': 20, 'n_estimators': 400}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	5005.05	0.82
Test	5740.90	0.75

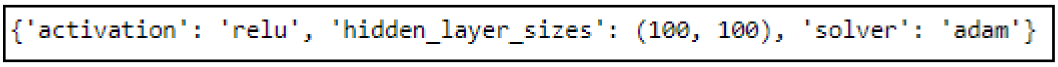
GridSearchCV for ANN regressor:

The given tuning parameters are below:



```
GridSearchCV(cv=3, estimator=MLPRegressor(max_iter=500, random_state=1),
             param_grid={'activation': ['tanh', 'relu'],
                         'hidden_layer_sizes': [500, (100, 100)],
                         'solver': ['sgd', 'adam']})
```

The best parameters after fitting the model are:



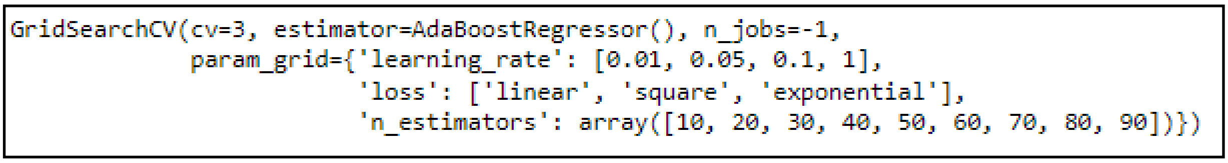
```
{'activation': 'relu', 'hidden_layer_sizes': (100, 100), 'solver': 'adam'}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	6270.43	0.71
Test	6075.72	0.72

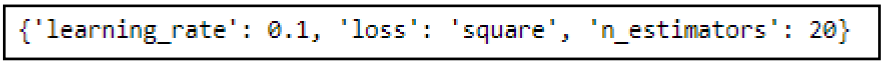
GridSearchCV on AdaBoosting:

The given tuning parameters are below:



```
GridSearchCV(cv=3, estimator=AdaBoostRegressor(), n_jobs=-1,
             param_grid={'learning_rate': [0.01, 0.05, 0.1, 1],
                         'loss': ['linear', 'square', 'exponential'],
                         'n_estimators': array([10, 20, 30, 40, 50, 60, 70, 80, 90])})
```

The best parameters after fitting the model are:



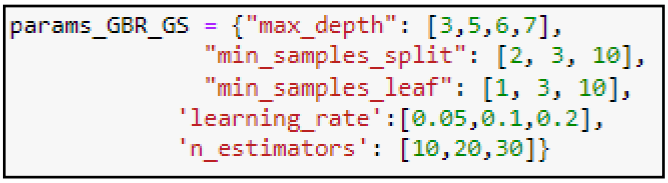
```
{'learning_rate': 0.1, 'loss': 'square', 'n_estimators': 20}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	6717.76	0.67
Test	6535.12	0.68

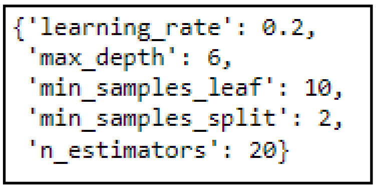
GridSearchCV on Gradient Boosting:

The given tuning parameters are below:



```
params_GBR_GS = {"max_depth": [3,5,6,7],
              "min_samples_split": [2, 3, 10],
              "min_samples_leaf": [1, 3, 10],
            'learning_rate':[0.05,0.1,0.2],
            'n_estimators': [10,20,30]}
```

The best parameters after fitting the model are:



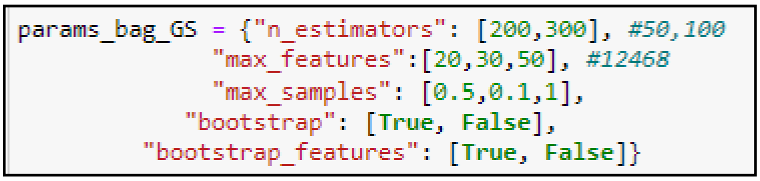
```
{'learning_rate': 0.2,
 'max_depth': 6,
 'min_samples_leaf': 10,
 'min_samples_split': 2,
 'n_estimators': 20}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	5535.96	0.77
Test	5730.60	0.75

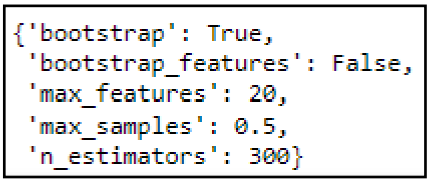
Bagging with GridSearchCV:

The given tuning parameters are below:



```
params_bag_GS = {"n_estimators": [200,300], #50,100
              "max_features":[20,30,50], #12468
              "max_samples": [0.5,0.1,1],
            "bootstrap": [True, False],
         "bootstrap_features": [True, False]}
```

The best parameters after fitting the model are:



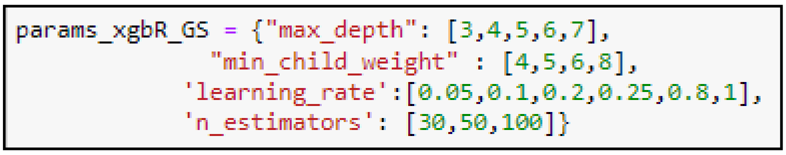
```
{'bootstrap': True,
 'bootstrap_features': False,
 'max_features': 20,
 'max_samples': 0.5,
 'n_estimators': 300}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	6185.21	0.72
Test	6707.60	0.66

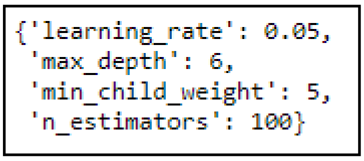
GridSearchCV on XGB Regressor:

The given tuning parameters are below:



```
params_xgbR_GS = {"max_depth": [3,4,5,6,7],
              "min_child_weight" : [4,5,6,8],
            'learning_rate':[0.05,0.1,0.2,0.25,0.8,1],
            'n_estimators': [30,50,100]}
```

The best parameters after fitting the model are:



```
{'learning_rate': 0.05,
 'max_depth': 6,
 'min_child_weight': 5,
 'n_estimators': 100}
```

Test Results after Tuning:

	RMSE	Accuracy Score
Train	5370.65	0.79
Test	5719.79	0.75

	Train RMSE	Test RMSE	Training Score	Test Score
Ridge Regression with GridSearchCV	6560.29	6398.49	0.68	0.69
Lasso Regression with GridSearchCV	6308.69	6110.29	0.71	0.72
Linear Regression with GridSearchCV	6307.43	6110.81	0.71	0.72
Decision Tree Regressor with GridSearchCV	5527.60	5933.77	0.77	0.73
Random Forest Regressor with GridSearchCV	5005.05	5740.90	0.82	0.75
ANN Regressor with GridSearchCV	6270.43	6075.72	0.71	0.72
AdaBoost Regressor withGridSearchCV	6717.76	6535.12	0.67	0.68
Gradient Boosting R egressor with GridSearchCV	5535.96	5730.60	0.77	0.75
Bagging Regressor with GridSearchCV	6185.21	6707.60	0.72	0.66
XGB Regressor with GridSearchCV	5370.65	5719.79	0.79	0.75

Table 5: Hypertuned models results after hypertuning

5. Model Validation:

Two metrics there are use to validate how well a model fits a dataset are the root mean squared error (RMSE) and the R-squared (R2), which are calculated as follows:

- **RMSE:** A metric that tells us how far apart the predicted values are from the observed valuesin a dataset, on average. The lower the RMSE, the better a model fits a dataset.

Formula 5: $$RMSE = \sqrt{\frac{(e_1^2+e_2^2+\cdots+e_n^2)}{n}}\text{, where } e_i = y_i - \widehat{y_\imath}$$

- **R-Square(R2):** A metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. This value ranges from 0 to 1. The higher the R2 value, the better a model fits a dataset.

Formula 6: $$\boldsymbol{R^2 = 1 - \frac{SSE}{SST}}$$

Where, $$\boldsymbol{SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \text{ and } SSE = \sum_{i=1}^{n}\left(Y_i - \widehat{Y_\imath}\right)^2}$$

Final Model Comparison:

	Train RMSE	Test RMSE	Training Score	Test Score
Ridge Regression	6307.43	6110.81	0.71	0.72
Lasso Regression	6307.43	6110.81	0.71	0.72
Linear Regression	6307.43	6110.81	0.71	0.72
Decision Tree Regressor	0.00	8231.95	1.00	0.49
Random Forest Regressor	2267.59	5918.08	0.96	0.74
ANN Regressor	6270.43	6075.72	0.71	0.72
AdaBoost Regressor	6942.13	6849.01	0.64	0.65
Gradient Boosting Regressor	5967.89	5819.52	0.74	0.74
Bagging Regressor	2684.79	6150.14	0.95	0.71
XGB Regressor	4380.68	5991.81	0.86	0.73
Ridge Regression with GridSearch CV	6560.29	6398.49	0.68	0.69
Lasso Regression with GridSearchCV	6308.69	6110.29	0.71	0.72
Linear Regression with GridSearchCV	6307.43	6110.81	0.71	0.72
Decision Tree Regressor with GridSearchCV	5527.60	5933.77	0.77	0.73
Random Forest Regressor with GridSearchCV	5005.05	5740.90	0.82	0.75
ANN Regressor with GridSearchCV	6270.43	6075.72	0.71	0.72
AdaBoost Regressor with				
GridSearchCV	6717.76	6535.12	0.67	0.68
Gradient BoostingRegressor with				
GridSearchCV	5535.96	5730.60	0.77	0.75

Table 6: Final Model Comparison

Comparing all the model performed, the most optimal model is Random Forest regressor. Random Forest regressor is giving 82% accuracy In Train & 75% in Test with RMSE value of Train as 5005.05 &Train as 5740.90.

Feature Importance of Random Forest:

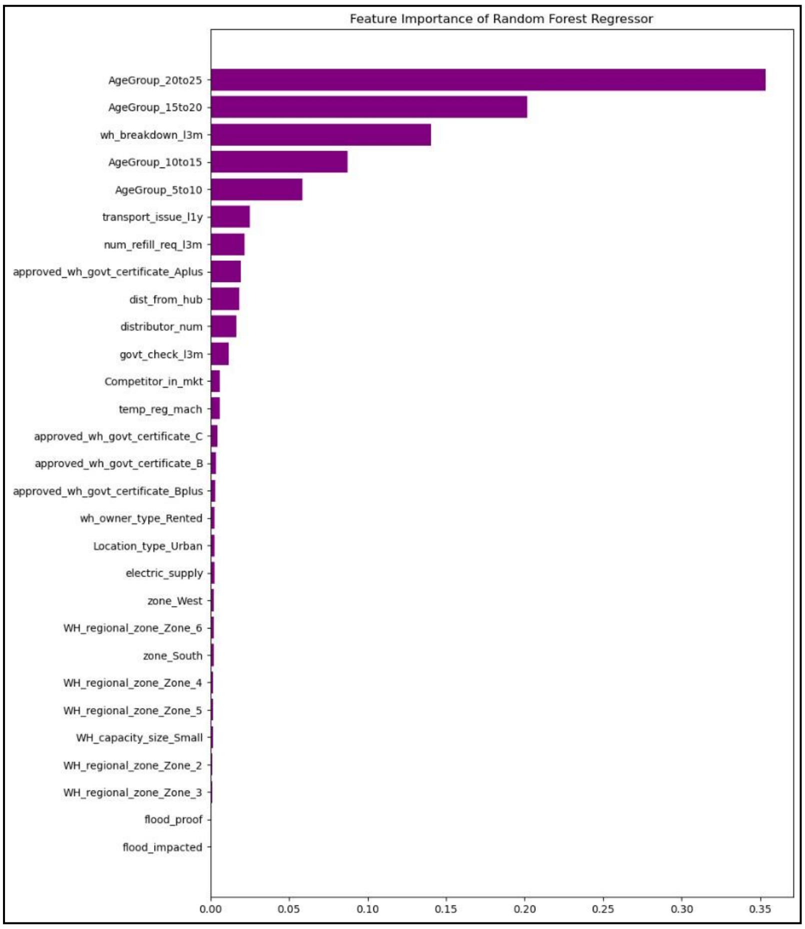


Figure 15: Feature Importance of Random Forest

These features have maximum effect on optimum weight (product_wg_ton):

1. Warehouse Breakdown in last 3 months.
2. Transport issue.
3. Refilling time in last 3 months.
4. Warehouse established year.

6. Final interpretation / recommendation:

Interpretation of the most optimum model and its implication on business.

Optimal model:

- Among all the models performed till now, XGB Regressor & Random Forest are performing well.
- We can find that Ridge, Lasso is not showing much improvement in the score and RMSE aftertuning and results of both are almost same.
- Decision Tree regressor has improved after performing model tuning.
- ANN regressor is giving the same values even after performing the hypertuning.
- We can observe that Gradient boost regressors has improved.

Model Insights:

- Since this is regression problem, we have tried out different regression models to confirm which performs well and gives the best accuracy.
- To handle overfitting, we performed hyperparameter tuning using GridSearchCV.
- Comparing all models, we have obtained the best results for Random Forest regressor → Random Forest regressor records 82% accuracy In Train & 75% in Test with RMSE value of Train as 4998.51 & Train as 5753.39
- The warehouse breakdowns due to both internal & external factors result on its inventory management & leading manufactures.
- Accident or Product stolen shall be another factor which can lead to optimum weight mismatch at the time of delivery, resulting in supply constraints
- Delay in stock refilling hampers reduced stock during high demand times
- Features that affect product_wg_ton which specifies the optimum weight of the product to be shipped are Warehouse that are established at least 5 years ago and its importance increases with the age of warehouse.

Implication on the business:

Based on the model outputs, the business should strive to improve across Strategic, Operational and tactical dimensions. These improvements have the potential to transform business, by supplychain being a revenue lever, beyond just cost savings.

1. Strategic dimension:

A. Improved strategic KPIs:

- Strategic KPIs such as Customer Service Level and improved fill rates needs to be improved using better demand planning and forecasting. Warehouse operations needs to be aligned to deliver inventory management and accurate inventory records. This helps to know whenever there is a refill required immediately and won't affect supply.

B. Improved strategic focus:

- East zone's product demand is on par with other zones, however with increased competition. However, the zone has fewer warehouses, retail outlets and distributors. Hence it is critical to focus on improving more focus on this zone, by enhancing the product fulfilment experience.

2. Operational dimension:

A. Warehouse efficiency:

- Set up a governing council that offers a clear strategy for functionality and efficiency, thus reducing Warehouse breakdown factors. The council's aim is to give directions and align thesupply chain strategy with the company's core goals. The council helps in removing barrierswithin the organization.
- Need to perform frequent audits upon warehouse operation standards. Use technology

to improve the supply chain. Review all the existing processes that are affecting the inventory management. Determine the areas where implementing technology could improve the processes.

B. Operational streamlining:

- Review policies and procedures to ensure efficiency and compliance. It also helps avoid bottlenecks in the supply chain, streamline operations and mitigate the risks of theft and fraud. Regular reviews help in identifying different risk elements and estimating their financial impact.

3. Tactical dimension:

A. Distribution:

- Distribution from hub is a key feature identified in impacting the business – hence improved logistics scheduling can play a key role for improvement in this area.