Feb 2022 | PG-DSBA-Online

SUPPLY CHAIN MANAGEMENT PROJECT REPORT

SUBMITTED BY
DEV TRIPATHI

Contents

Section 1 1

Introduction 1

Problem Statement 1

Need of this analysis 1

Understanding the business opportunity 1

Data Report 3

Data Dictionary 3

Data Collection 3

About the Dataset 4

Exploratory Data Analysis 6

Univariate analysis 6

Bivariate Analysis 18

Data Preprocessing 21

Clustering 23

Section 2 27

Modelling 27

Train and Test split 27

Linear Regression Model 27

Random Forest Model 30
Support Vector Regressor Model 31
XGBOOST Model 32
Comparison b/w models 34
Recommendations 35

List of Figures

Figure 1: Data Dictionary 3
Figure 2: Feature information and null count for these feature (in %) 4
Figure 3: Boxplot and distribution plot for 'retail_shop_num' variable 6
Figure 4: Boxplot and distribution plot for 'distributor_num' variable 7
Figure 5: Boxplot and distribution plot for 'dist_from_hub' variable 7
Figure 6: Boxplot and distribution plot for 'workers_num' variable 8
Figure 7: Boxplot and distribution plot for 'storage_issue_reported_l3m' variable 8
Figure 8: Boxplot and distribution plot for 'product_wg_ton' variable 9
Figure 9: Boxplot and distribution plot for 'age_wh' variable 9
Figure 10: Countplots for categorical variables 17
Figure 11: Correlation plot for numeric features present in the dataset 18
Figure 12: Zone vs. Warehouse Breakdown with location type 19
Figure 13: Values of 'wh_est_year' grouped by various categorical columns 21
Figure 14: Silhouette width calculated based on identified clusters 23
Figure 15: Visualization of formed clusters 24
Figure 16: Performance metrics for LR model (1st case) 28
Figure 17: Results obtained from LR 28
Figure 18: Performance metrics on the train (left) and test (right) set 29
Figure 19: Plots for residual analysis of LR model 29
Figure 20: Performance metrics obtained from RF model (1st case) 30
Figure 21: Performance metrics obtained from RF model (2nd case) 30
Figure 22: Performance metrics for tuned RF model (3rd case) 31
Figure 23: Performance metrics obtained from SVR model (case 1) 31
Figure 24: Performance metrics for SVR model (2nd case) 32

Figure 25: Performance metrics for SVR model (3rd case)..... 32
Figure 26: Performance metrics for XGBoost (1st case)..... 32
Figure 27: Performance metrics for XGBoost (2nd case)..... 33
Figure 28: Performance metrics for XGBoost model (3rd case) 33

Section 1

Introduction

Problem Statement

An FMCG company has entered started manufacturing instant noodles two years back. The higher management in the company has noticed a mismatch between supply and demand. Where the demand is high, supply is pretty low, and where the demand is low, supply is pretty high. Since this can cause a considerable amount of inventory cost loss, higher management has decided to optimize the supply chain. The product quantity being supplied to each and every warehouse established in the entire country is to be optimized as per the demand for the that particular location.

Need of this analysis

Supply chain optimization is one of the keys to business success, especially in the FMCG sector, because the competition has increased many folds. The FMCG companies have to make their products available to the right customer at the right time in the right quantity; otherwise, the consumers generally buy similar products available in the market. Also, companies like this one, which has recently entered manufacturing the product, need to focus on the supply and demand as their consumer base is comparatively small.

Understanding the business opportunity

The food processing industry is expected to at a rapid pace. According to industry estimates, the food processing industry accounts for nearly 30% of the total food market in India. Furthermore, the total food production in India is estimated to double in the next 10 years. Following are the factors which are expected to fuel the growth in this sector:

- Increasing spending on health and nutritional foods

- An increasing number of nuclear families and working women
- Changing lifestyle
- Functional foods, fresh or processed foods
- Organized retail and private label penetration
- Changing demographics and rising disposable incomes

Data Report

Data Dictionary

The dataset contains a total of 24 features. The description of these variables is given in Figure 1.

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

Figure 1: Data Dictionary

Data Collection

To solve this particular problem, the data required must have been collected from various departments such as the HR department, production department, logistics department etc., present

in the concerned company. In our case, company managed to provide us data for warehouses present in different zone and regions. Though by looking at the data we can say that the company has put appreciable amount of efforts to maintain their records as most the entries present in the dataset, were observed to be very less to no missing data at all.

About the Dataset

```
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)
```

```
Null values present in each feature (in %):

Ware_house_ID                  0.000
WH_Manager_ID                  0.000
Location_type                  0.000
WH_capacity_size               0.000
zone                           0.000
WH_regional_zone               0.000
num_refill_req_l3m             0.000
transport_issue_l1y            0.000
Competitor_in_mkt              0.000
retail_shop_num                0.000
wh_owner_type                  0.000
distributor_num                0.000
flood_impacted                 0.000
flood_proof                    0.000
electric_supply                0.000
dist_from_hub                  0.000
workers_num                    3.960
wh_est_year                   47.524
storage_issue_reported_l3m     0.000
temp_reg_mach                  0.000
approved_wh_govt_certificate   3.632
wh_breakdown_l3m               0.000
govt_check_l3m                 0.000
product_wg_ton                 0.000
```

Figure 2: Feature information and null count for these feature (in %)

Observations:

- The dataset contains 24 variables and 25000 entries for these variables.
- There 8 features are of object datatype, 2 features are of float datatype and 14 features are integer datatype.
- Only 3 features are having missing values which are **'wh_est_year' (47.5%), 'workers_num' (4%), and 'approved_wh_certificate' (3.632%)**.
- Though the 'wh_est_year' should have been removed as it contains more than 40% values as missing values, we chose to keep it after imputing it with a suitable value.

- Also, we have imputed the missing values present in the dataset with median values for **'workers_num', 'approved_wh_certificate'** features and by mode value for **'wh_est_year'.**
- For further analysis, the **'wh_est_year'** feature was converted to **'age_wh'**, representing the warehouse's age at the **present date (2023).**
- Also, the **'zone'** and **'WH_regional_zone'** were concatenated to become one single variable **'Zone'**.

Exploratory Data Analysis

Before performing EDA, we dropped two variables warehouse ID and warehouse manager ID as these would not help to understand or get insights about the data.

Univariate analysis

Continuous Features:

1. **'retail_shop_num'**

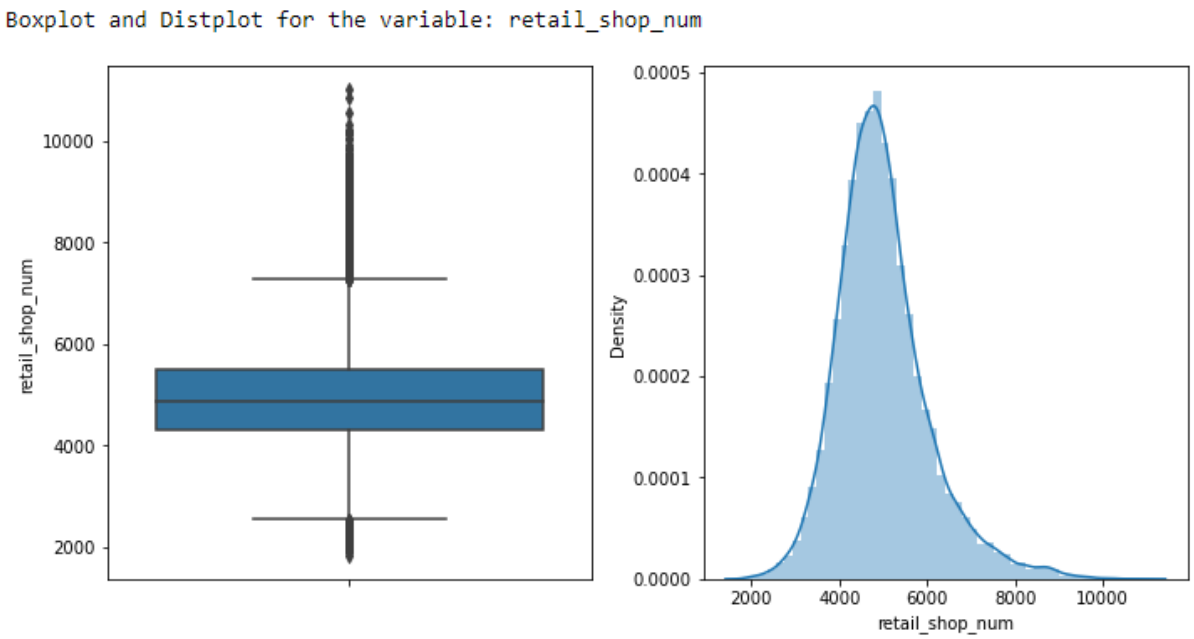
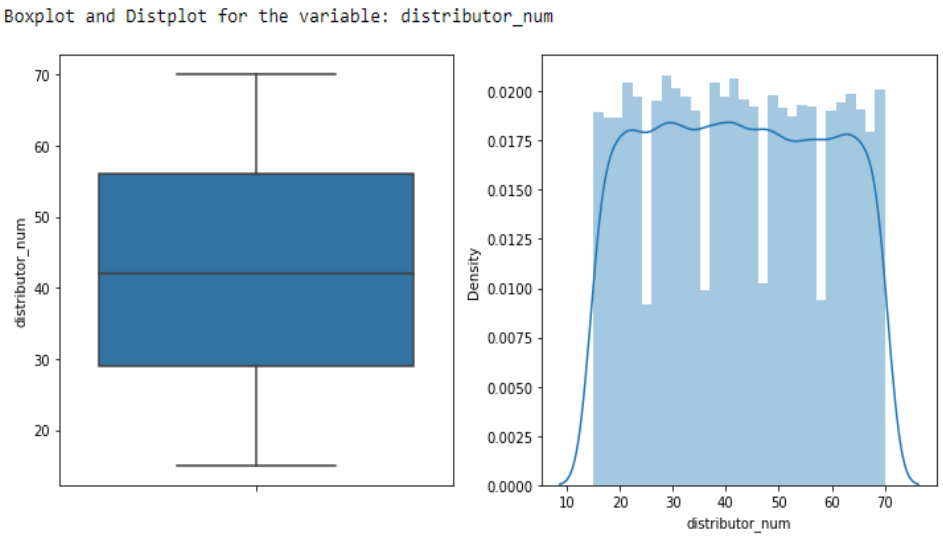


Figure 3: Boxplot and distribution plot for 'retail_shop_num' variable

Observations:

- From the above plot, we can say that the distribution is right-skewed.
- The Median is around 5000
- Outliers are present in the data for this feature

2. distributor_num

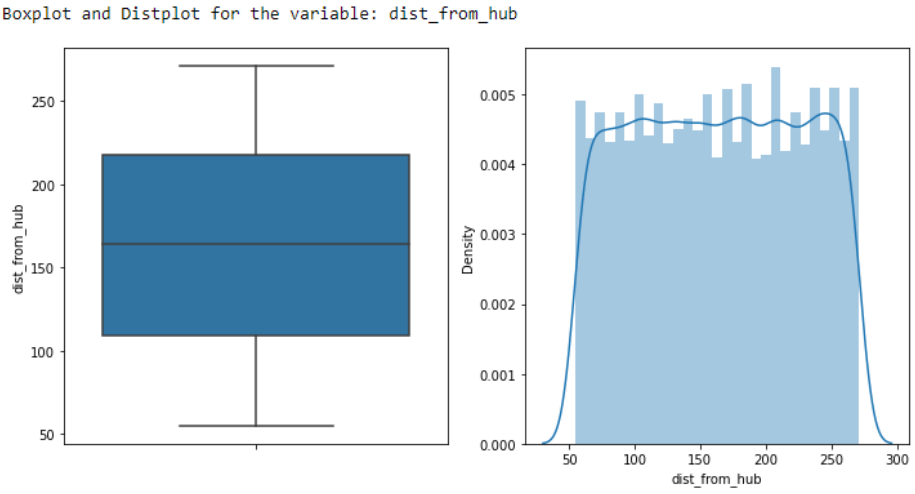


Observations:

The median value is 42. There are no outliers present in the data for this feature. The data has very low to nil skewness.

Figure 4: Boxplot and distribution plot for 'distributor_num' variable

3. dist_from_hub

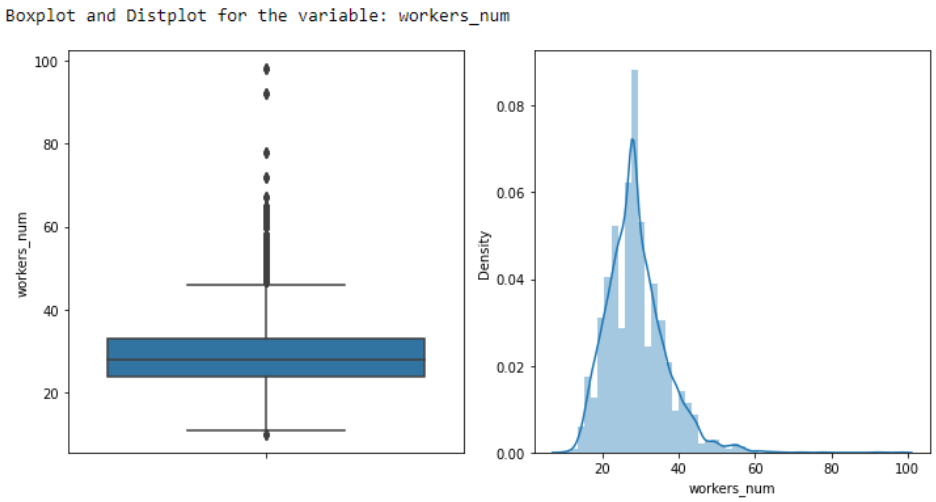


Observations:

We can say from the above plot that the distribution has very low skewness. The Median is around 165. Outliers are not present in the data for this feature.

Figure 5: Boxplot and distribution plot for 'dist_from_hub' variable

4. workers_num

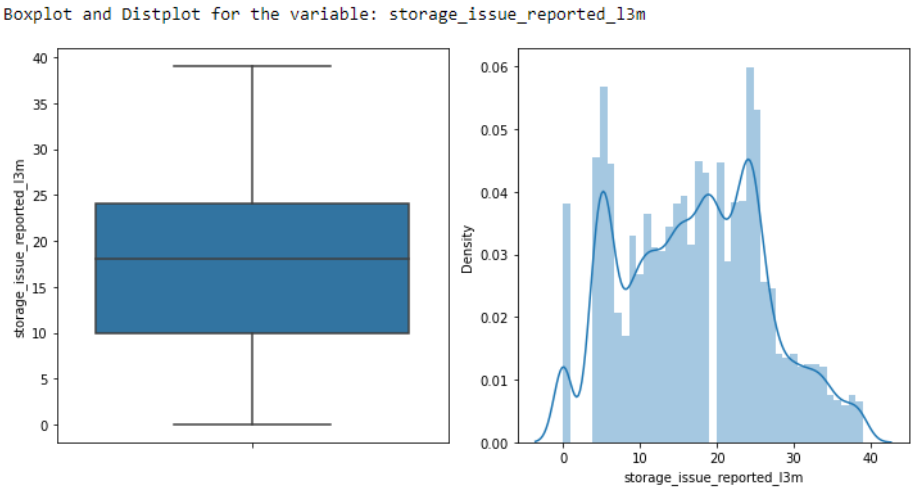


Observations:

From the above plot, we can say that the distribution is right-skewed. The Median is around 29. Outliers are present in the data for this feature.

Figure 6: Boxplot and distribution plot for 'workers_num' variable

5. storage_issues_reported_l3m



Observations:

From the plot, we can say that the distribution is slightly right-skewed. The Median is around 18. Outliers are not present in the data for this feature.

Figure 7: Boxplot and distribution plot for 'storage_issue_reported_l3m' variable

6. product_wg_ton

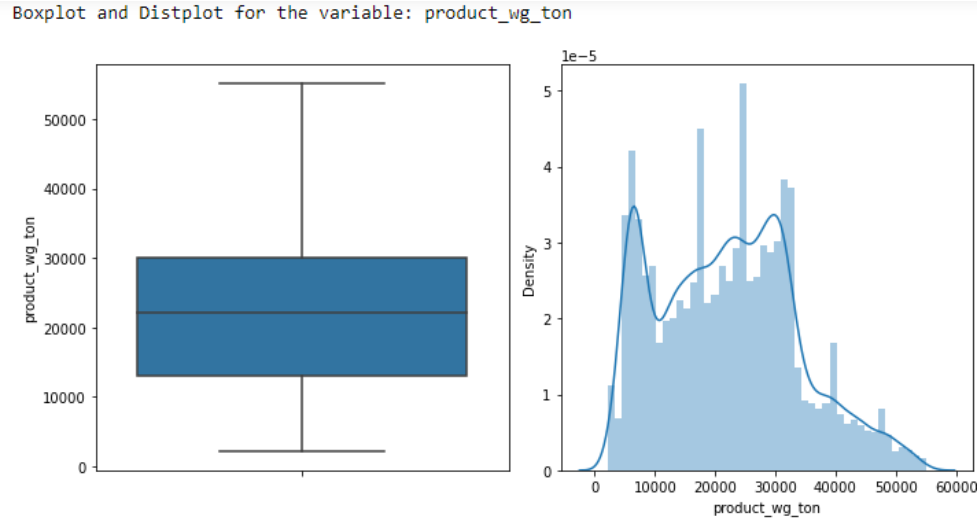


Figure 8: Boxplot and distribution plot for 'product_wg_ton' variable

Observations:

From the plot, we can say that the distribution is right-skewed. The Median is around 25000. Outliers are not present in the data for this feature.

7. age_wh

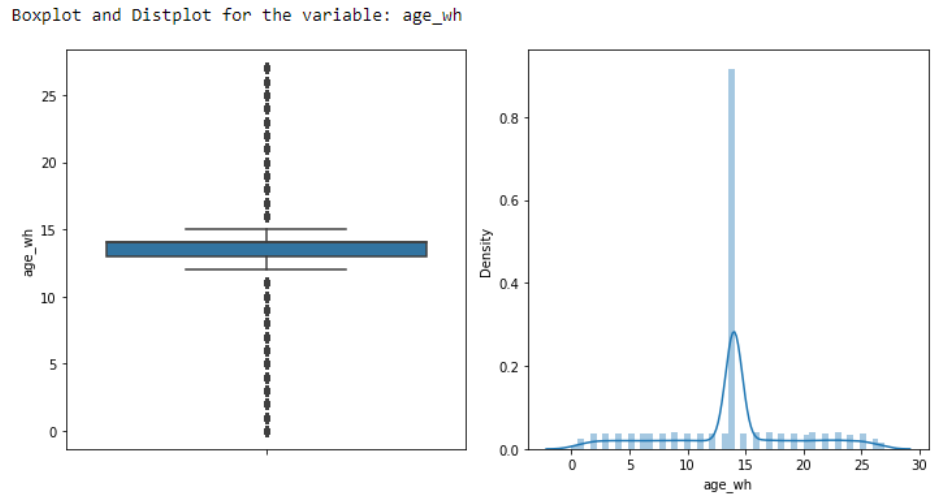


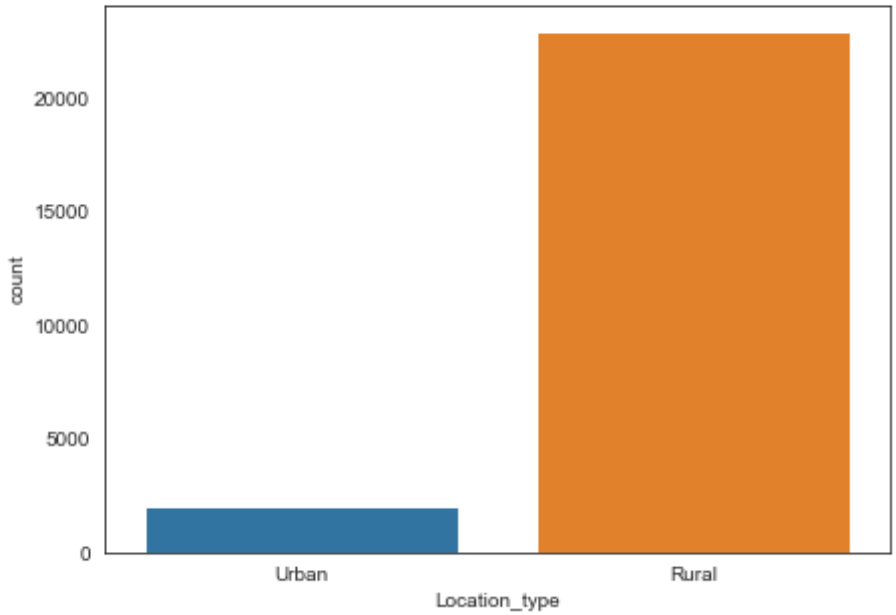
Figure 9: Boxplot and distribution plot for 'age_wh' variable

Observations:

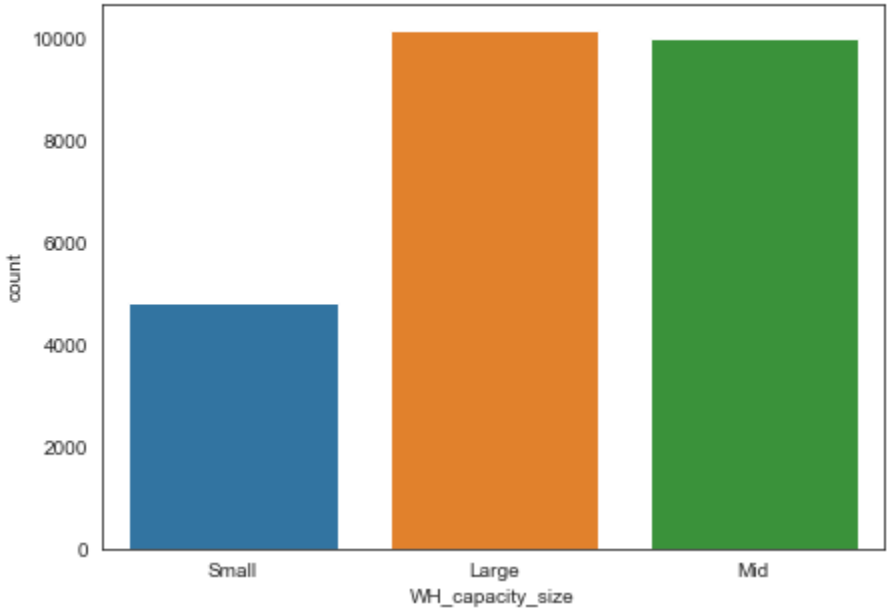
Since we have imputed the 47% missing values and then calculated 'age_wh', the insights are not relevant here.

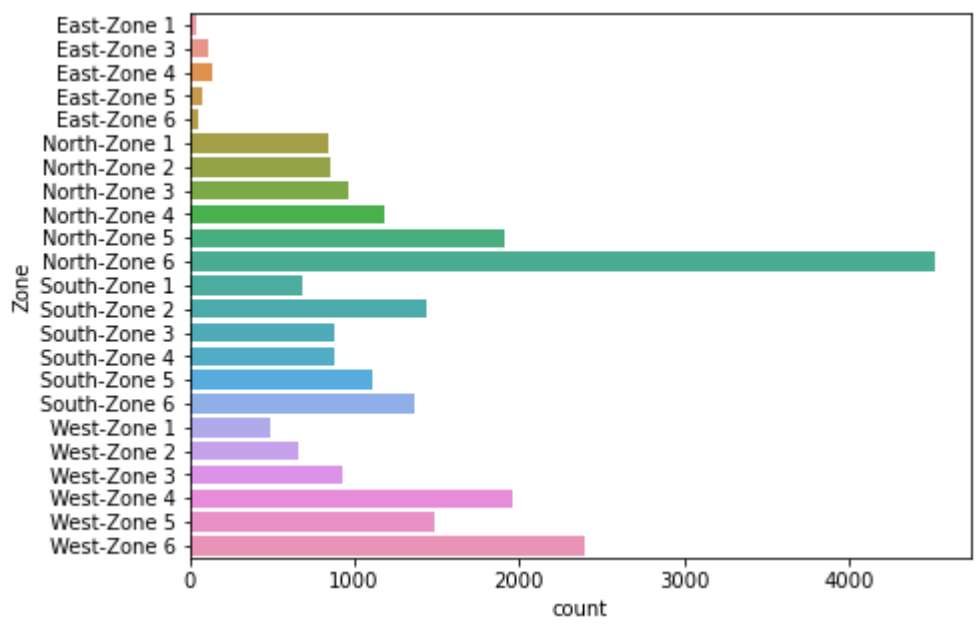
Categorical Features:

Countplot for variable: Location_type

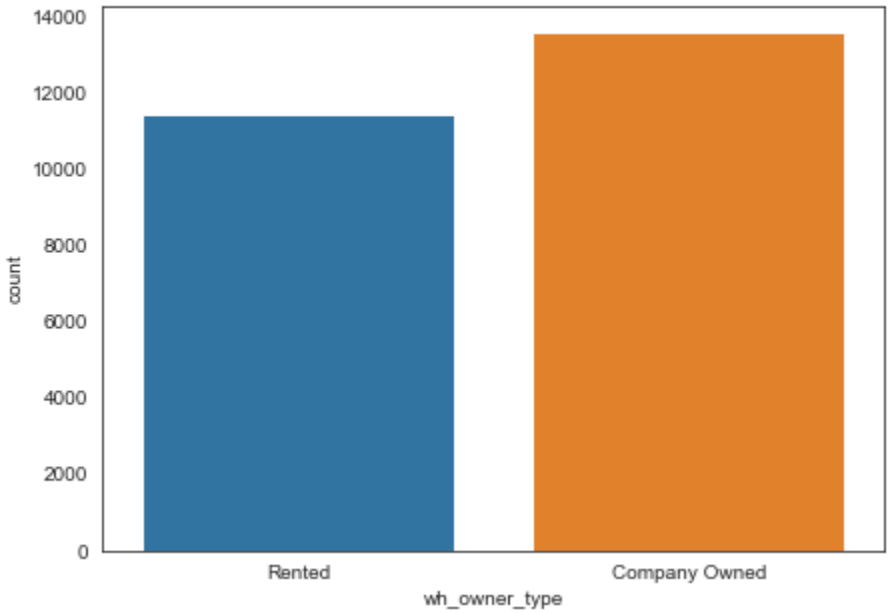


Countplot for variable: WH_capacity_size

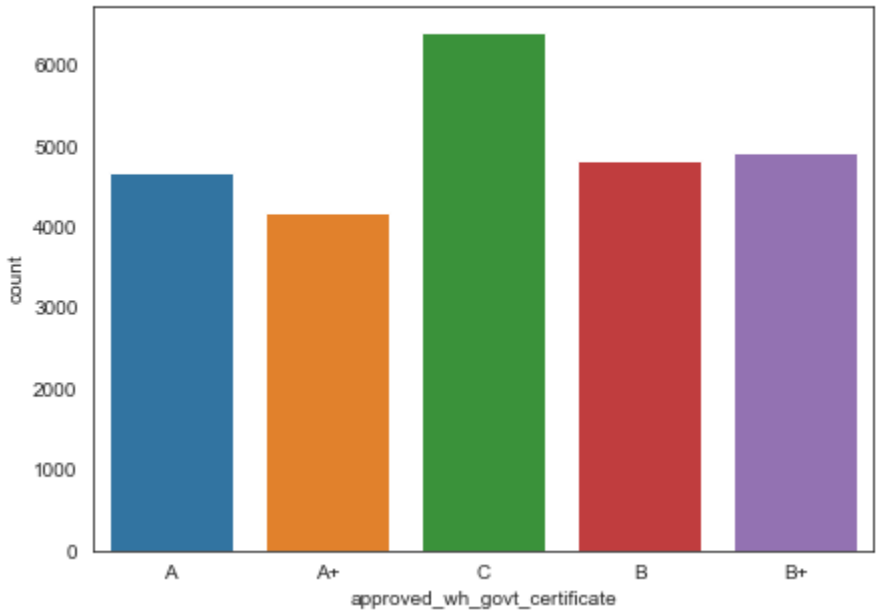




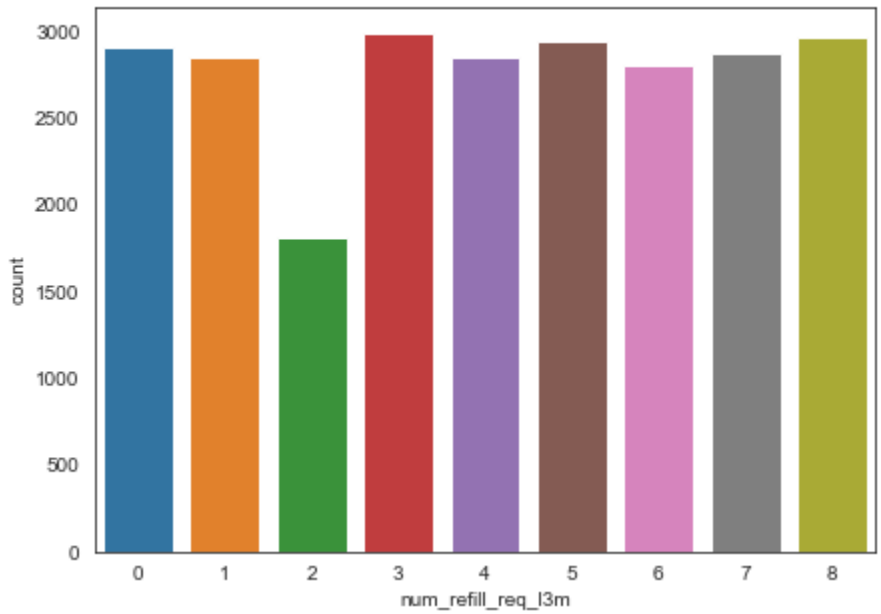
Countplot for variable: wh_owner_type



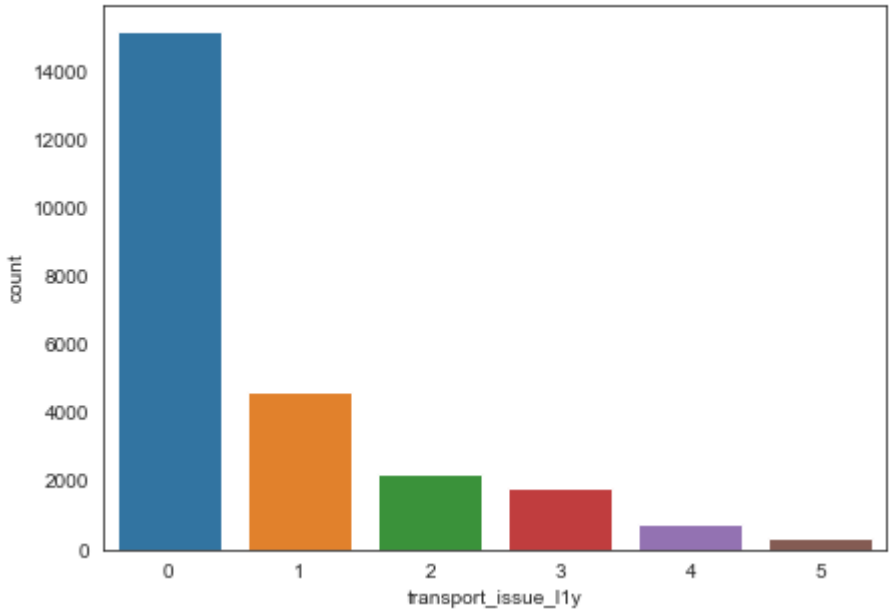
Countplot for variable: approved_wh_govt_certificate



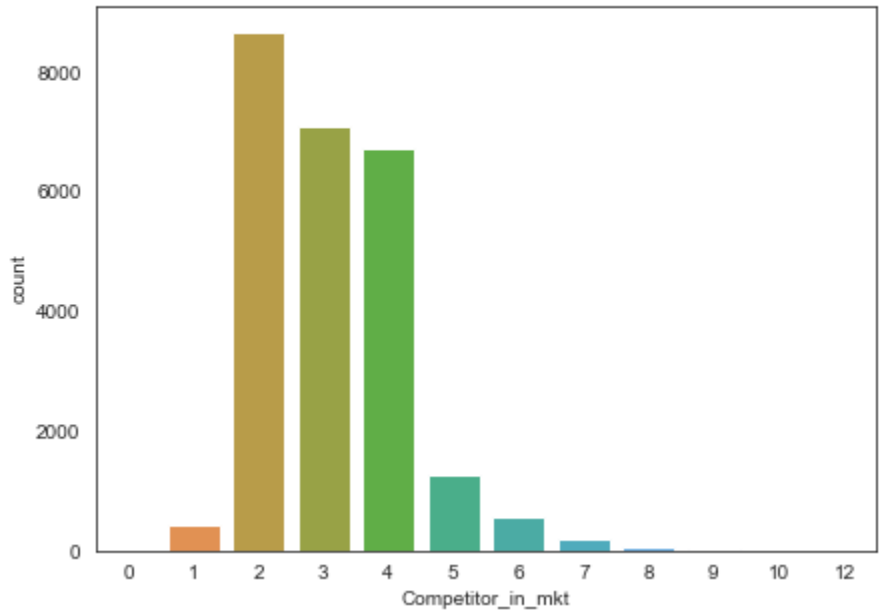
Countplot for variable: num_refill_req_l3m



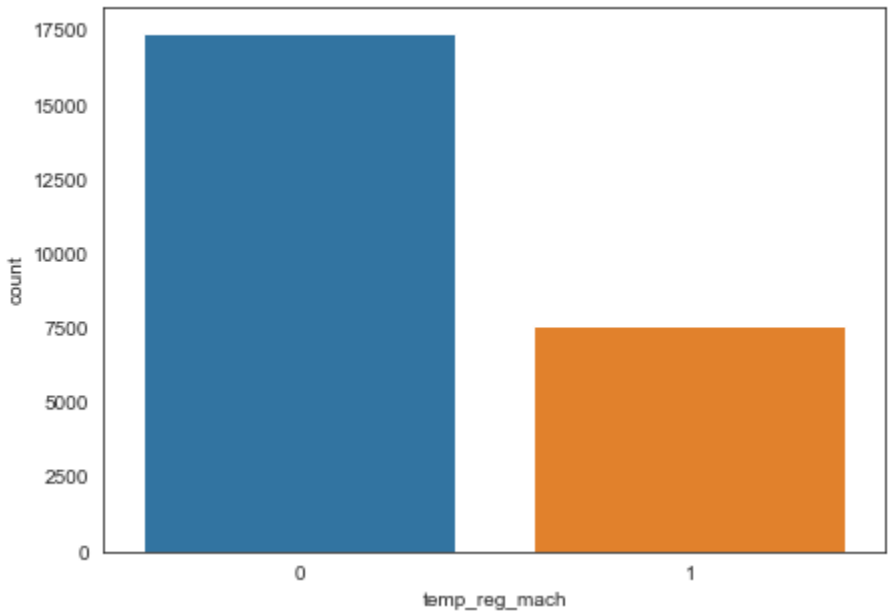
Countplot for variable: transport_issue_l1y



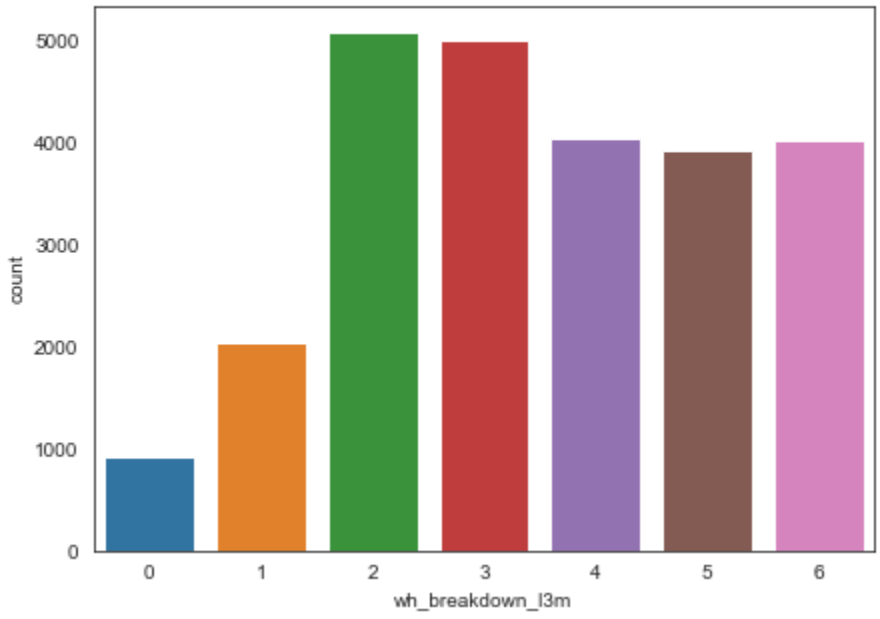
Countplot for variable: Competitor_in_mkt



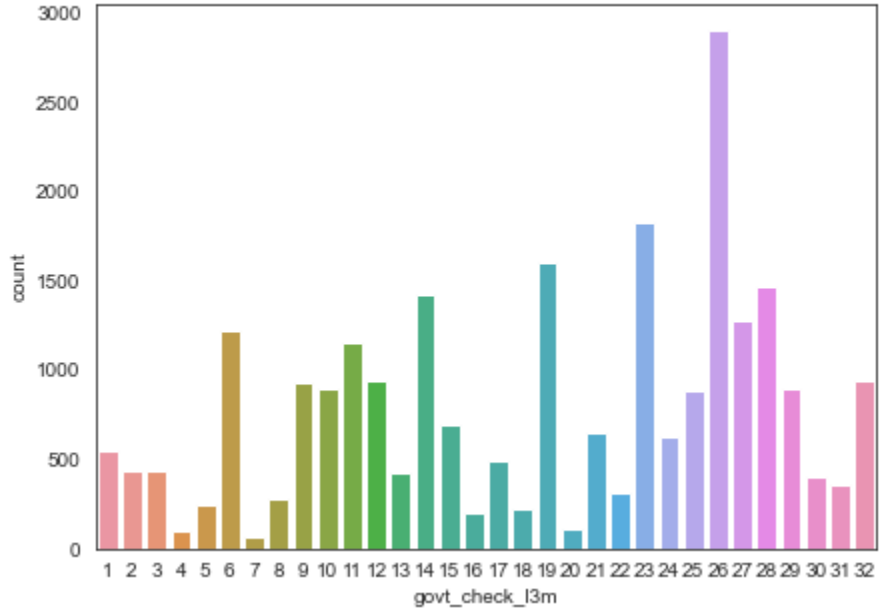
Countplot for variable: temp_reg_mach



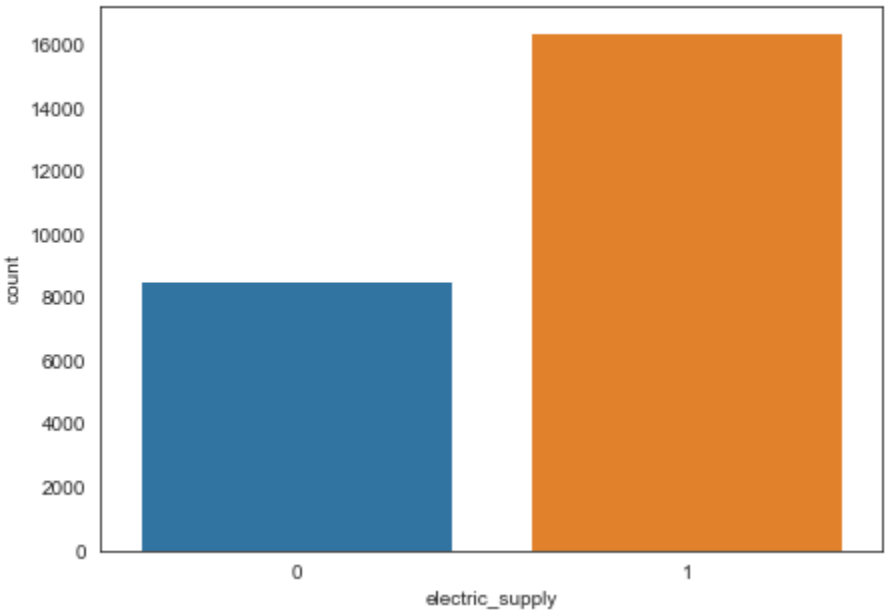
Countplot for variable: wh_breakdown_l3m



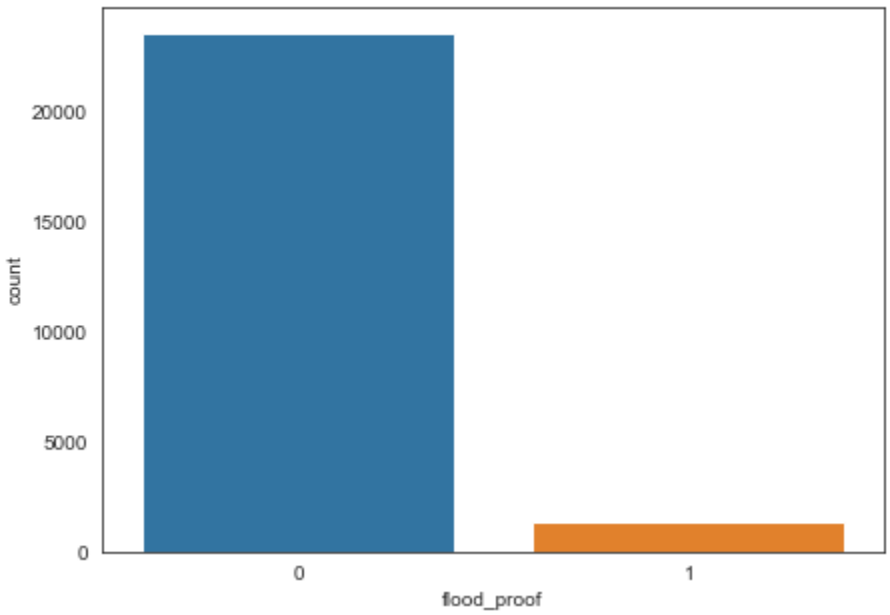
Countplot for variable: govt_check_l3m



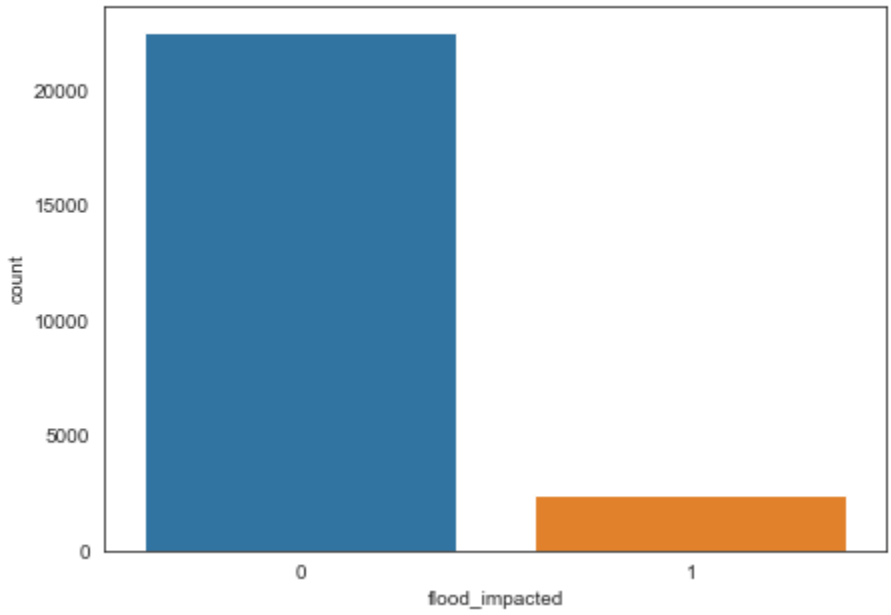
Countplot for variable: electric_supply



Countplot for variable: flood_proof



Countplot for variable: flood_impacted



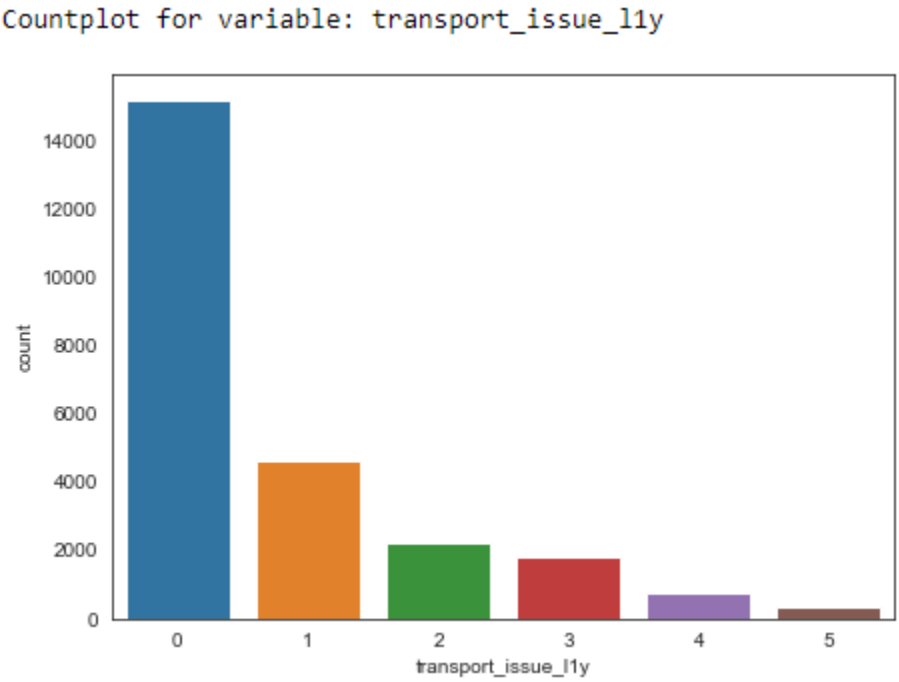


Figure 10: Countplots for categorical variables

Observations:

- Most of the warehouses are located in **'Rural'** area.
- The **'small'** warehouses are less than half in numbers compared to **'large'** or **'mid'** size warehouses.
- The highest number of warehouses are located in **'North-Zone 6**' followed by **'West-Zone 6'**
- The number of refills varies from 0 to 8. Strangely, warehouses that required two refills in the past three months are significantly less than all other values
- Transport issues in last one year are having zero as mode value which is good for business.
- For most of the entries, the competitors in the market are between 2 to 4.

- Out of 25000 warehouses, 17500 warehouses do not have temperature regulatory machines.
- The number of times warehouse breakdown happened ranges between 2 to 6 in the past 3 months.
- The number of times government checks happened is having mode value equal to 26. The variable ranges from 1 to as high as 32
- More than 16000 warehouses are having electric supply
- More than 23000 warehouses are flood proof
- More than 23000 warehouses are flood impacted.

Bivariate Analysis

Correlation plot:

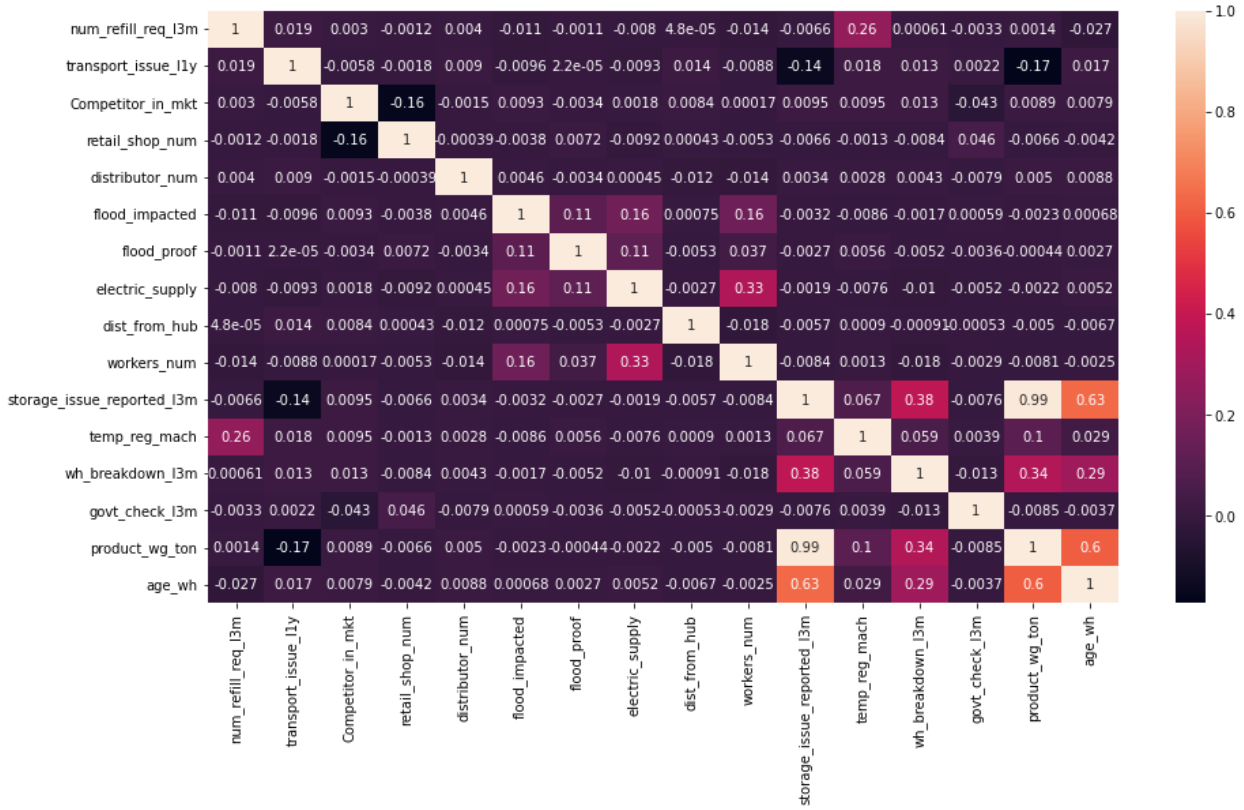


Figure 11: Correlation plot for numeric features present in the dataset

Observations:

- Most of the features have very little to no correlation at all
- Our target feature **'product_wg_ton'** is having very high correlation **(0.99)** with **'storage_issues_reported_l3m'** and moderate correlation with **'age_wh'**
- **'age_wh'** and **'product_wg_ton'** are also having a high correlation **(0.63)**

Zone vs. Warehouse Breakdown with location type as a filter:

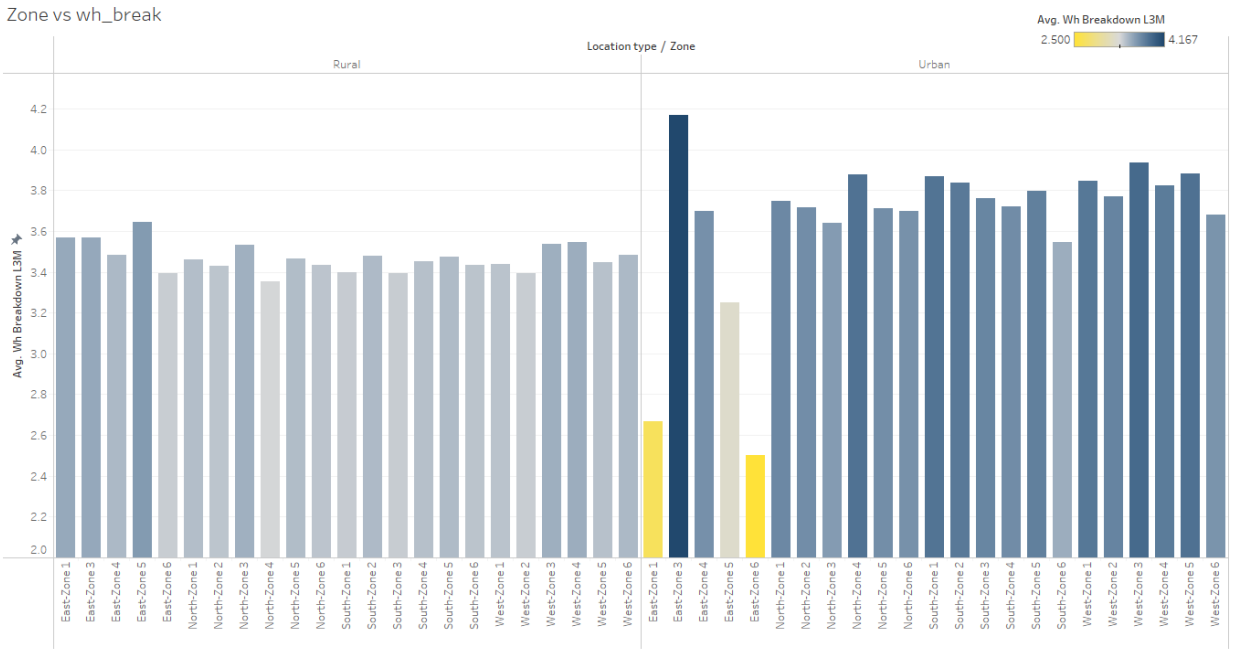


Figure 12: Zone vs. Warehouse Breakdown with location type

Observations:

- For the Urban and East zone 3, warehouses are having maximum average number of breakdowns

From the Figure 12, it is obvious that the number of breakdowns for Urban area is more than rural area.

Data Preprocessing

Before proceeding to modelling, data cleaning and preprocessing is required. Also, as mentioned earlier some features are having null values. Imputation of null values is also to be done using suitable values. In this analysis, we took following steps to eradicate such anomalies from the data:

1. Firstly, imputation of null values present in the 'wh_est_year' column was done. For this purpose the values were grouped by categorical column to observe any pattern present.

zone	WH_regional_zone	WH_capacity_size	
East	Zone 1	Small	2011.0
	Zone 3	Mid	2008.0
	Zone 4	Mid	2007.0
	Zone 5	Large	2008.0
	Zone 6	Small	2008.0
North	Zone 1	Small	2009.0
	Zone 2	Mid	2010.0
	Zone 3	Mid	2010.0
	Zone 4	Mid	2009.0
	Zone 5	Large	2009.0
	Zone 6	Large	2009.0
		Small	2009.0
South	Zone 1	Small	2010.0
	Zone 2	Mid	2009.0
	Zone 3	Mid	2010.0
	Zone 4	Mid	2010.0
	Zone 5	Large	2010.0
	Zone 6	Large	2008.0
		Small	2010.0
West	Zone 1	Small	2009.0
	Zone 2	Mid	2009.0
	Zone 3	Mid	2009.0
	Zone 4	Mid	2009.0
	Zone 5	Large	2009.0
	Zone 6	Large	2009.0
		Small	2009.0

Figure 13: Values of 'wh_est_year' grouped by various categorical columns

Now, the null values were imputed by 2009 in 'wh_est_year' as for most of the values are 2009 which are present in the data (Figure 13).

2. Further, the null values in ‘approved_wh_govt_certificate’ and ‘workers_num’ were imputed with mode values of their respective features.
3. From the data we observed that there is a region (named as ‘zone’) and sub-region (named as ‘wh_regional_zone’). Since these are complementary to each other for locating any warehouse, we created a new column named as ‘Zone’ in which we concatenated the region and sub-region. e.g. if zone is ‘North’ and sub-region is ‘Zone 1’ for a warehouse then the new value ‘Zone’ is ‘North-Zone 1’ for that particular warehouse.
4. Lastly, the ‘zone’ and ‘wh_regional_zone’ variables were dropped from the dataset.
5. For some numerical variables, outliers were present but for this analysis, the outliers trement was not perfomed since there are only 7 numerical features overall.

Clustering

Finally, K- Means clustering was used to detect some patterns or clusters from the dataset. Although the Elbow Curve did not show any significant drop in within sum of squares values when plotted against the number of clusters, we assumed the number of clusters as 3. Boxplot of Silhouette width calculated for these clusters was generated (Figure 13):

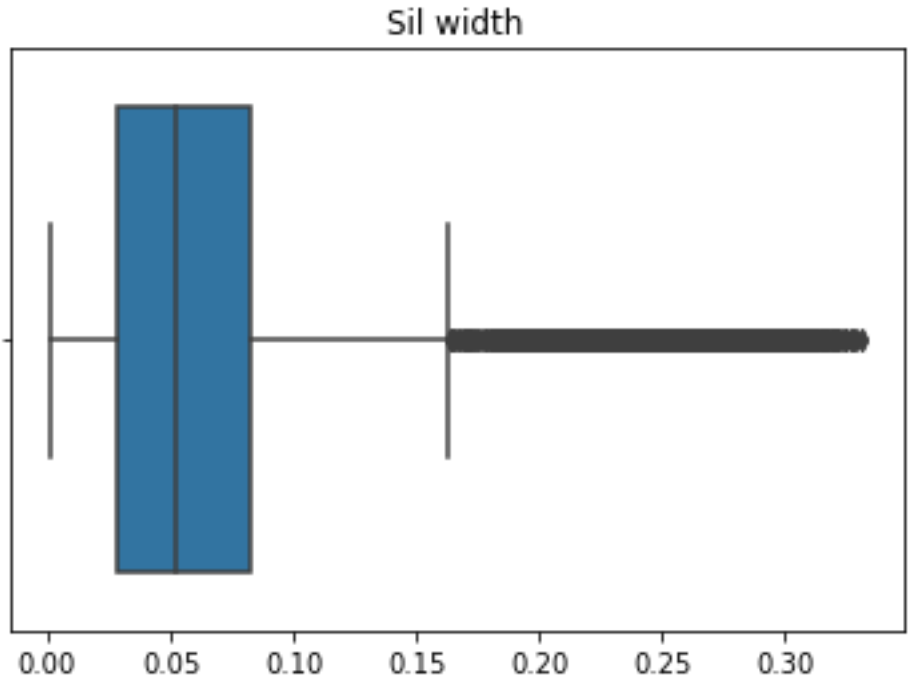


Figure 14: Silhouette width calculated based on identified clusters

Silhouette score for the formed clusters was coming out to be +0.064. Now, for visualization of formed clusters using Principal Components Analysis:

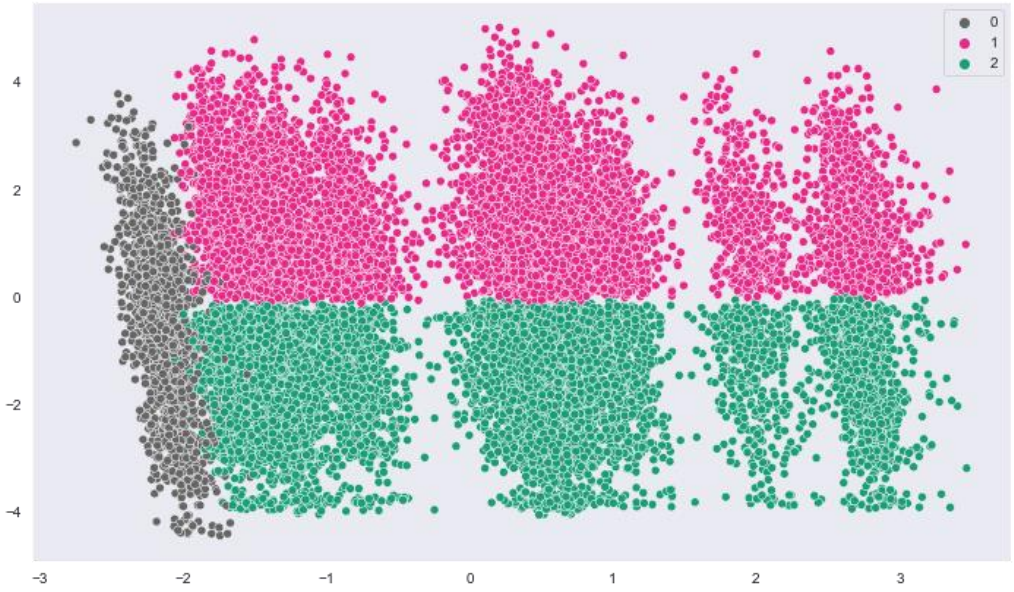


Figure 15: Visualization of formed clusters

Now, let's have a look at the differences between the clusters identified:

	WH_capacity_size	storage_issue_reported_l3m	wh_breakdown_l3m	product_wg_ton	Zone	labels
count	12896	12896	12896	12896	12896	12896
unique	3	NaN	NaN	NaN	21	NaN
top	Mid	NaN	NaN	NaN	North - Zone 6	NaN
freq	5605	NaN	NaN	NaN	2490	NaN
mean	NaN	23.777605	4.11616	30440.7	NaN	2
std	NaN	5.828665	1.416405	7982.53	NaN	0
min	NaN	7	1	10081	NaN	2
25%	NaN	20	3	25067	NaN	2
50%	NaN	24	4	29130	NaN	2
75%	NaN	27	5	34102	NaN	2
max	NaN	39	6	55151	NaN	2

Table 1: Statistical description for cluster 1

	WH_capacity_size	storage_issue_reported_l3m	wh_breakdown_l3m	product_wg_ton	Zone	labels
count	10541	10541	10541	10541	10541	10541
unique	3	NaN	NaN	NaN	21	NaN
top	Mid	NaN	NaN	NaN	North - Zone 6	NaN
freq	4415	NaN	NaN	NaN	2029	NaN
mean	NaN	9.042785	2.705246	11952.5	NaN	1
std	NaN	5.016138	1.671482	5736.29	NaN	0
min	NaN	0	0	2065	NaN	1
25%	NaN	5	1	7067	NaN	1
50%	NaN	9	2	11149	NaN	1
75%	NaN	13	4	16133	NaN	1
max	NaN	23	6	30139	NaN	1

Table 2: Statistical description of cluster 2

	WH_capacity_size	storage_issue_reported_l3m	wh_breakdown_l3m	product_wg_ton	Zone	labels
count	1563	1563	1563	1563	1563	1563
unique	1	NaN	NaN	NaN	2	NaN
top	Large	NaN	NaN	NaN	West - Zone 5	NaN
freq	1563	NaN	NaN	NaN	1489	NaN
mean	NaN	16.829814	3.488804	21759.9	NaN	0
std	NaN	9.308235	1.709459	11789.9	NaN	0
min	NaN	0	0	3058	NaN	0
25%	NaN	9	2	12121.5	NaN	0
50%	NaN	17	3	22058	NaN	0
75%	NaN	24	5	29147.5	NaN	0
max	NaN	39	6	55111	NaN	0

Table 3: Statistical description of cluster 3

Observations:

- There is total 12896 warehouses identified as cluster-1, 10841 warehouses identified as cluster-2 and 1563 warehouses identified as cluster-3
- If we compare the identified cluster-1 with cluster-2, the mean values of **'strorage_issues_reported_l3m'**, **'wh_breakdown_l3m'** and **'product_wg_ton'** are significantly different.
- On the other hand, cluster-3 has the mean values of the same variables in between the values obtained for cluster-1 and cluster-2.
- Almost all the warehouses categorized as cluster-3 are located in **'West-Zone 5' (1589 out of 1683).**
- All the warehouses categorized as cluster-3 have **'Large'** warehouse capacity.

Section 2

Modelling

In this section, we will be building various models and validating those models on the test set. But firstly, we need to prepare the data before feeding it to the models.

Train and Test split

Now, we have to split the data in train and test set separately to train and then check the performance of models on the test set. For this analysis, we have chosen the ratio of train and test set as 70:30, which means 70 percent of the data will be fed to the train set, and the rest goes for the test set.

Now, it was observed in the fore mentioned chapters that the correlation of `storage_issue_l3m' is highly correlated with the target variable, i.e., 'product_wg_ton'. Hence, we have tried random forest, SVR, and XGBoost models for three iterations.

1. Including the 'storage_issue_l3m' variable
2. Excluding the 'storage_issue_l3m' variable
3. Excluding the 'stoage_issue_l3m' variable and a better-tuned model (after searching for hyperparameters)

Linear Regression Model

Now, in the modelling step, we started with the linear regression model. Being simple and interpretable in nature, we can get better insights compared to other models.

LR model for 1st case

For the 1st case, as we have included the highly correlated variable `storage_issue_l3m', we obtain very good performance for both train and test sets. The performance metrics obtained are mentioned below:

```
Performance for LR model for Train set (1st case)
RMSE:  1740.2991514344594
MAE:  1279.883461579977
R2 score:  0.9775666793733301
```

```
Performance for LR model for Test set (1st case)
RMSE:  1763.4388193113027
MAE:  1297.5018018598219
R2 score:  0.9768100313550326
```

Figure 16: Performance metrics for LR model (1st case)

As we can observe, the LR model seems to perform very well on the train and test set. The model fits well on the data set as the performance is similar on the train and test set.

LR model for 2nd case

In this case, we have built the linear regression model after dropping the highly correlated feature 'storage_issue_l3m'. Since there are so many statistically insignificant variables, we have applied the Recursive Feature Elimination technique to choose the best five features among these variables, with a p-value less than 0.05. The results obtained are given below:

OLS Regression Results

Dep. Variable:	product_wg_ton	**R-squared:**	0.451
Model:	OLS	**Adj. R-squared:**	0.450
Method:	Least Squares	**F-statistic:**	2392.
Date:	Sun, 13 Feb 2022	**Prob (F-statistic):**	0.00
Time:	11:15:45	**Log-Likelihood:**	-1.8340e+05
No. Observations:	17500	**AIC:**	3.668e+05
Df Residuals:	17493	**BIC:**	3.669e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.21e+04	65.114	339.392	0.000	2.2e+04	2.22e+04
transport_issue_l1y	-2199.0461	65.076	-33.792	0.000	-2326.601	-2071.491
temp_reg_mach	592.6455	68.205	8.689	0.000	458.957	726.334
approved_wh_govt_certificate	1319.2547	69.443	18.998	0.000	1183.139	1455.370
wh_breakdown_l3m	1880.1800	68.981	27.256	0.000	1744.970	2015.390
age_wh	6315.0191	68.382	92.349	0.000	6180.984	6449.054
Urban	423.4812	65.099	6.505	0.000	295.881	551.081

Omnibus:	169.426	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	174.166
Skew:	0.242	**Prob(JB):**	1.51e-38
Kurtosis:	3.066	**Cond. No.**	1.48

Figure 17: Results obtained from LR

As it can be clearly observed that the performance of the model is very poor from the summary of LR model itself. The model in mathematical terms can be given as:

'product_wg_ton' = -2199.0461* **'transport_issue_l1y'** +592.6455* **'temp_reg_mach'** +1319.2547* **'approved_wh_govt_certificate'** +1880.1800* **'wh_breakdown_l3m'** +423.4812* **'Urban'** + 6315.0191* **'age_wh'** + 2.21*e^4

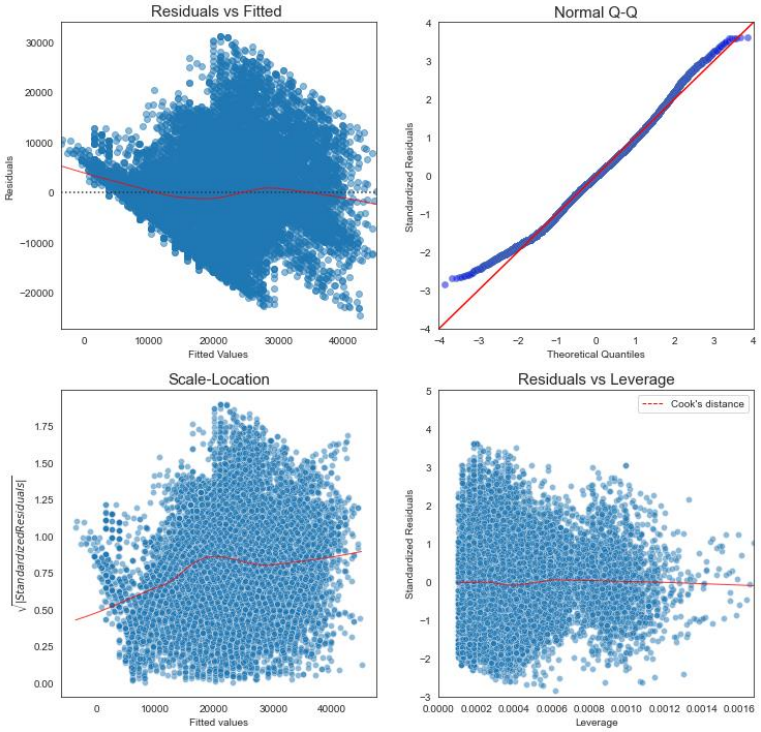
The performance metrics for this model are given below:

```
RMSE:  8611.68322255074
MAE:  6784.7879309357595
R2 score:  0.4506845187381755
```

```
RMSE:  8591.098662065093
MAE:  6785.172357166803
R2 score:  0.44960208393109735
```

Figure 18: Performance metrics on the train (left) and test (right) set

Now, although the model is not performing well, we can have a look at the residual plots to verify the assumptions of the linear regression.



As we can observe from figure 18, there is no pattern observed in the residuals, and also the residuals are following the normal distribution also. Hence, we can say that the assumptions are fairly true for this analysis.

Figure 19: Plots for residual analysis of LR model

Random Forest Model

Let's try a random forest model further to improve the performance of train and test sets.

RF model 1st case

For 1st case, the performance of the random forest model was found to be excellent as we have not dropped the `storage_issue_l3m' feature. The result obtained on the train and test set is given in figure 19

```
Performance for RF model for Train set (1st case)
RMSE:  342.65117932736536
MAE:  249.4382845714286
R2 score:  0.9991303385716575
```

```
Performance for RF model for Test set (1st case)
RMSE:  902.1360777982813
MAE:  667.3164386666665
R2 score:  0.9939309115031695
```

Figure 20: Performance metrics obtained from RF model (1st case)

RF model 2nd case

For the 2nd case, as expected, the performance of the model dropped significantly. Also, the model is overfitted on the train set, which can be observed from figure 20

```
Performance for RF model for Train set (2nd case)
RMSE:  2972.4769124544073
MAE:  2253.3870302857144
R2 score:  0.934554057440512
```

```
Performance for RF model for Test set (2nd case)
RMSE:  7884.133890480931
MAE:  6030.469424000001
R2 score:  0.5364598578094543
```

Figure 21: Performance metrics obtained from RF model (2nd case)

RF model 3rd case

For the 3rd case, we have tried to search for best hyperparameters using GridsearchCV method. The best hyperparameters obtained after GridsearchCV method are given as

{ 'max_depth' : [7],

'min_samples_leaf' : [3],

'min_samples_split' : [30],

'n_estimators' : [100]

}

The performance metrics obtained after fitting the model with parameters obtained from GridsearchCV are mentioned in figure 21.

```
Performance for RF model for Train set (3rd case)
RMSE:  7691.626470827885
MAE:  5881.630319509403
R2 score:  0.561790168037396
```

```
Performance for RF model for Test set (3rd case)
RMSE:  7796.901397681611
MAE:  5979.351298163209
R2 score:  0.5466606143650952
```

Figure 22: Performance metrics for tuned RF model (3rd case)

As we can observe, the model performance for both train and test sets is poor, but these are comparable for both sets with slight overfitting.

Support Vector Regressor Model

We tried building SVR models as well for all the fore-mentioned three cases description of which is given below:

SVR model case 1

For our dataset, the RBF kernel didn't seem to perform well. Hence, we have picked a linear kernel, which was giving fairly good results. The performance of the model was found to be good, which can be observed from the performance metrics mentioned in figure 22

```
Performance for SVR model for Train set (1st case)
RMSE:  2728.826038132791
MAE:  2059.2397483775426
R2 score:  0.944843405053282
```

```
Performance for SVR model for Test set (1st case)
RMSE:  2745.981298606326
MAE:  2073.3034044565966
R2 score:  0.943769166888346
```

Figure 23: Performance metrics obtained from SVR model (case 1)

SVR model case 2

Since we have dropped the highly correlated feature from the dataset in case 2, the performance of the SVR model has also dropped significantly. The model performance was even poor than the LR model (2nd case)

```
Performance for SVR model for Train set (2nd case)
RMSE:  8917.942088752596
MAE:  7006.604459285236
R2 score:  0.4109189635386459
```

```
Performance for SVR model for Test set (2nd case)
RMSE:  8885.690078985055
MAE:  6988.947476844104
R2 score:  0.411208284349314
```

Figure 24: Performance metrics for SVR model (2nd case)

SVR model case 3

Even after trying to search for hyperparameters, the performance does not seem to improve much. The performance metrics obtained for this model are given in Figure 24.

```
Performance for SVR model for Train set (3rd case)
RMSE:  8651.030378388137
MAE:  6757.336244518564
R2 score:  0.4456533584078861
```

```
Performance for SVR model for Test set (3rd case)
RMSE:  8640.089484997083
MAE:  6767.289624046715
R2 score:  0.44330688690892683
```

Figure 25: Performance metrics for SVR model (3rd case)

XGBOOST Model

Finally, we tried the extreme gradient boosting algorithms. The base value for all the cases was 0.5, and gradient boosted decision trees were used as the base model.

XGBoost Model Case 1

The performance of this model seems to be very good as the highly correlated feature is present in the dataset. The performance of this model was comparable in both train and test, as mentioned in figure 25.

```
Performance for XGBoost model for Train set (1st case)
RMSE:  613.0972806869889
MAE:  466.26698404017856
R2 score:  0.9972157737328323
```

```
Performance for XGBoost model for Test set (1st case)
RMSE:  883.77882197013
MAE:  652.4115538736979
R2 score:  0.9941753940654985
```

Figure 26: Performance metrics for XGBoost (1st case)

XGBoost Model Case 2

The performance of this model seems to be very poor as the highly correlated feature is absent in the dataset. The performance of this model is good for the train set, but it came out to be bad for the test set, which directly means that the model is overfitted on the test set.

```
Performance for XGBoost model for Train set (2nd case)
RMSE:  5903.952142813172
MAE:  4463.193529165867
R2 score:  0.7418147642862238
```

```
Performance for XGBoost model for Test set (2nd case)
RMSE:  8085.879328288185
MAE:  6251.144921369299
R2 score:  0.5124334764661822
```

Figure 27: Performance metrics for XGBoost (2nd case)

XGBoost Model Case 3

For the 3rd case, we have tried to search for best hyperparameters using GridsearchCV method. The best hyperparameters obtained after GridsearchCV method are given as

{ 'n_estimators' : [100],

'max_depth' : [3],

'min_child_weight' : [5],

'gamma' : [0],

'learning_rate' : [0.1] }

The performance of this model was comparable in both train and test, as mentioned in figure 27.

```
Performance for XGBoost model for Train set (3rd case)
RMSE: 7687.85416609965
MAE:  5910.950099546596
R2 score: 0.5622198965639169
```

```
Performance for XGBoost model for Test set (3rd case)
RMSE: 7744.525291558584
MAE:  5962.276751790365
R2 score: 0.5527308207853063
```

Figure 28: Performance metrics for XGBoost model (3rd case)

As we can observe, the model performance for both train and test sets is poor, but these are comparable for both sets with slight overfitting.

Comparison b/w models

Finally, if we compare all the models based on $RMSE, MAE\ and\ R^2\ scores$, we can say that the ensemble methods seem to perform best on both, i.e., Case 1 and Case 3. Though the values for XGBoost models seem to be more promising, the performance is quite similar on both train and test sets.

Model	Case	RMSE		MAE		R2 Score	
		Train	Test	Train	Test	Train	Test
Linear Regression	**1**	1740.299	1763.438	1279.83	1297.501	0.9775	0.976
	2	8611.683	8591.098	6784.787	6785.172	0.45	0.449
Random Forest	**1**	342.651	902.136	249.438	667.316	0.999	0.993
	2	2972.476	7884.133	2253.387	6030.469	0.934	0.536
	3	7691.626	7796.901	5881.63	5979.351	0.561	0.546
Support Vecto Regressor	**1**	2728.826	2745.981	2059.239	2073.303	0.944	0.943
	2	8917.942	8885.69	7006.604	6988.947	0.41	0.411
	3	8651.03	8640.89	6757.336	6767.289	0.445	0.443
Xgboost	**1**	613.097	883.778	466.266	652.411	0.997	0.994
	2	5903.952	8085.879	4463.193	6251.144	0.771	0.512
	3	7687.884	7744.525	5910.95	5962.276	0.562	0.552

Table 4: Comparison of various models trained based on performance metrics

Recommendations

Based on the analysis performed here, we have given following recommendations:

1. For almost all the models feature importance of the variables:
    a. **Age_wh**
    b. **Wh_breakdown_l3m**
    c. **Transport_issues_l3m**
    d. **Approved_wh_govt_certificate**
    e. **Location_type**

were coming out to be higher other than the highly correlated feature.

2. The above mentioned feature suggests that the issues related to the warehouse significantly impact the product weight to be shipped
3. Since ‘age_wh’ can play significant role in making good predictions, the client should provide accurate age of warehouses to get better results.
4. From EDA it was observed that the ‘East-Zone 3’ is having highest Average number of breakdown in last 3 months. The company should put efforts to minimize these as warehouse breakdown significantly affects the supply chain.
5. More significant features needs to be added to the dataset as the performance for almost all the models were found to be unsatifactoy after dropping the highly correlated feature.
6. The highly correlated feature ‘storage_issues_reported_l3m’ should be checked again as the high correlation causes suspicion. In case the data in this feature is found to be correct then it is recommended to use this feature clubbed with above mentioned features only. The model trained by using these variables only (above mentioned XGboost model for best performance), will be performing very well for all the practical purposes.