Feb 2021 | PG-DSBA-Online

SUPPLY CHAIN MANAGEMENT PROJECT REPORT

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

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'**

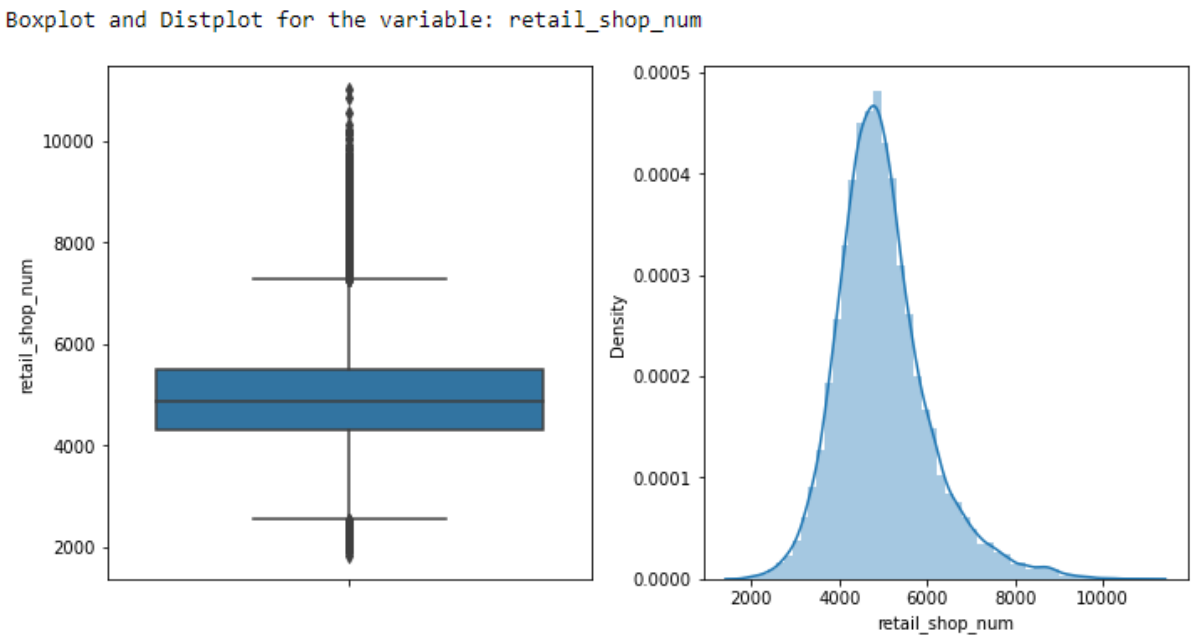
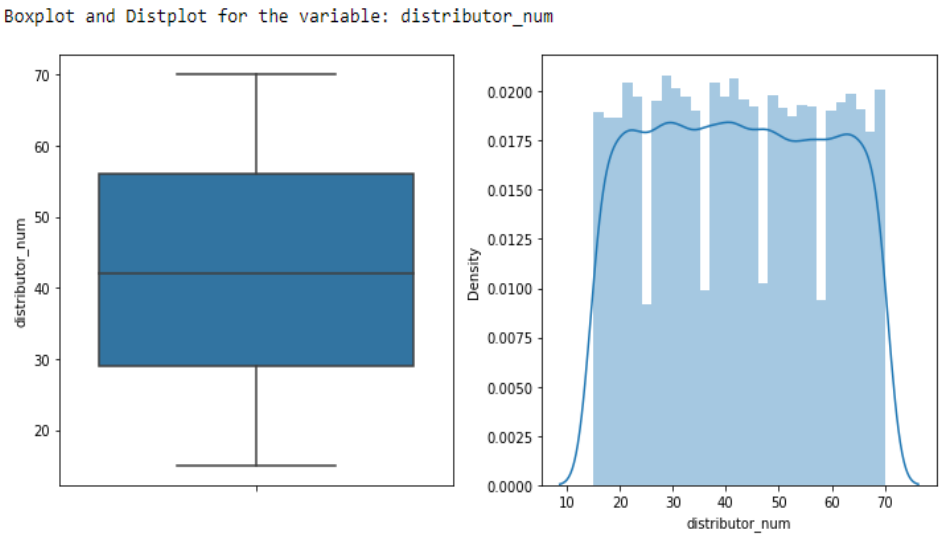


Figure 3: Boxplot and distribution plot for 'retail_shop_num' variable

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

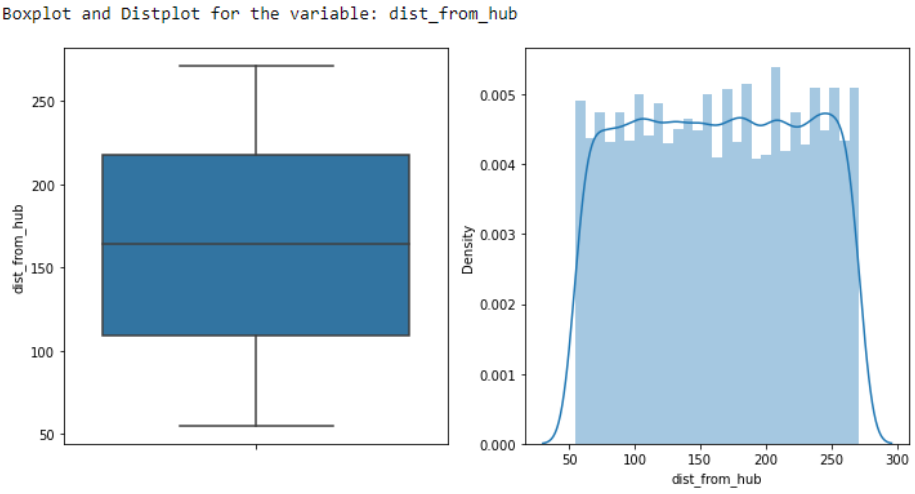


Observations:

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

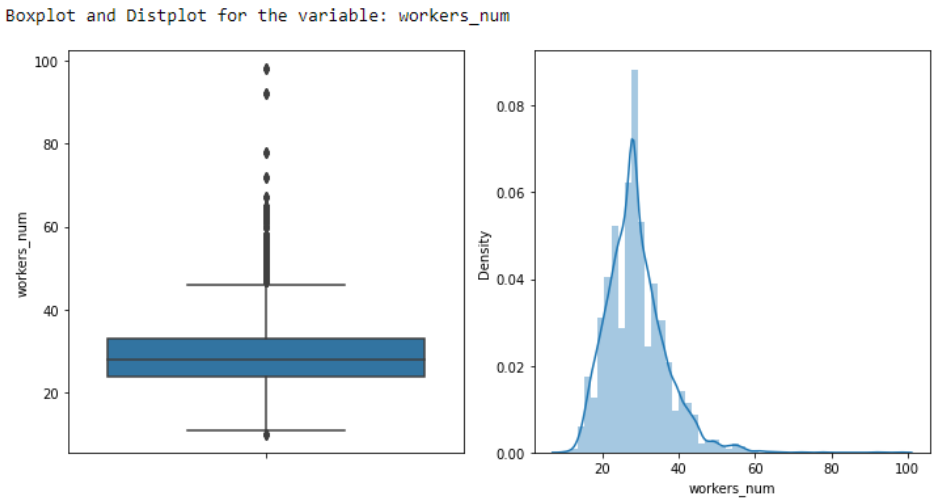


Observations:

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

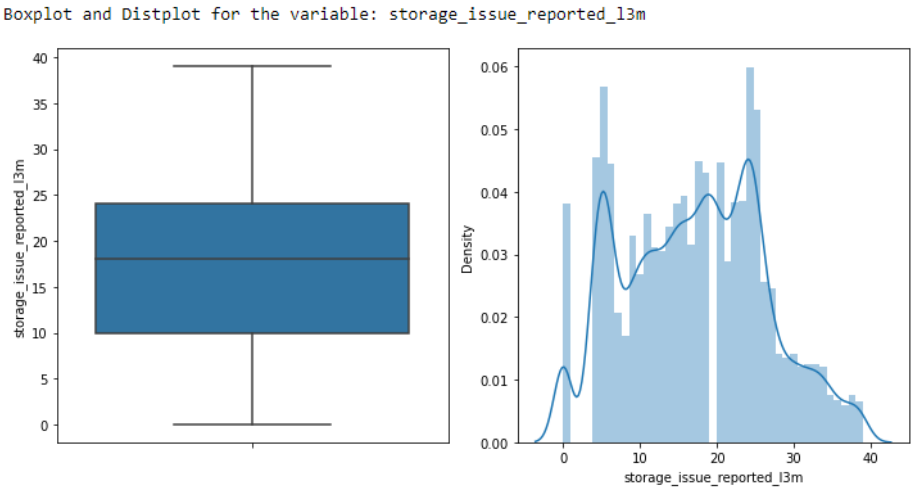


Observations:

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

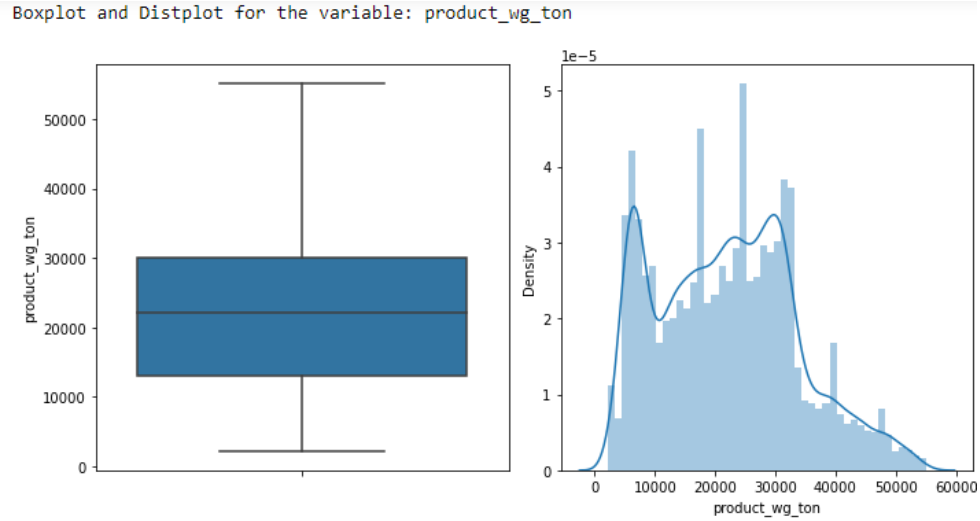


Figure 8: Boxplot and distribution plot for 'product_wg_ton' variable

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

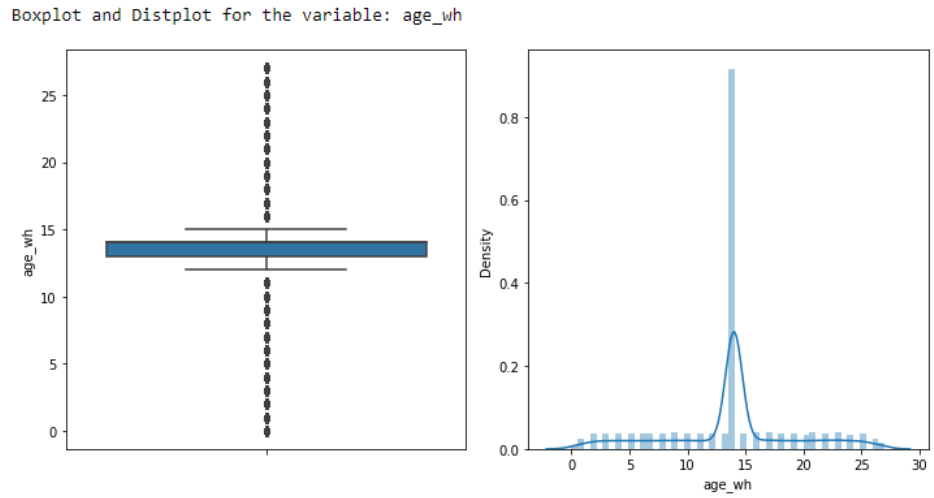


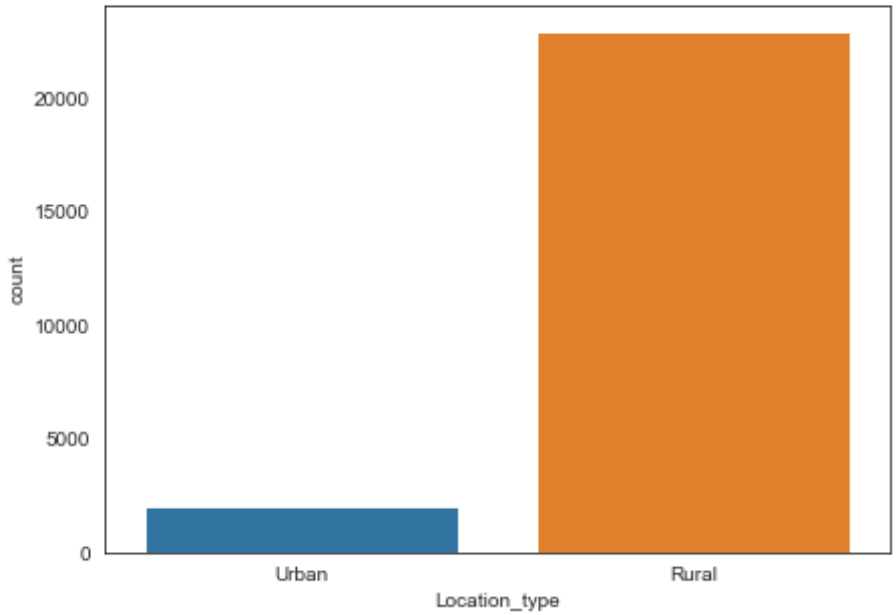
Figure 9: Boxplot and distribution plot for 'age_wh' variable

Observations:

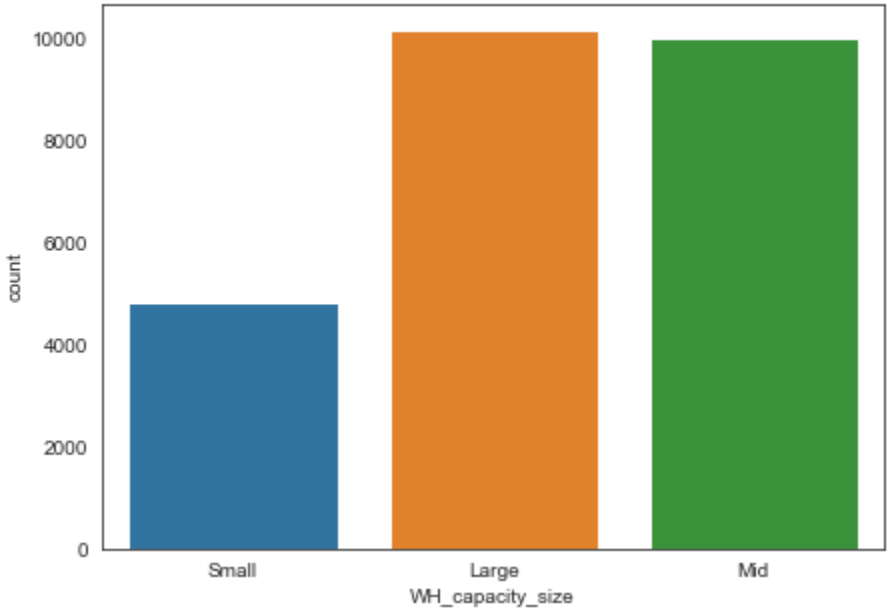
Since we have imputed the 47% missing values and then calculated 'age_wh', the insights are not relevant here.

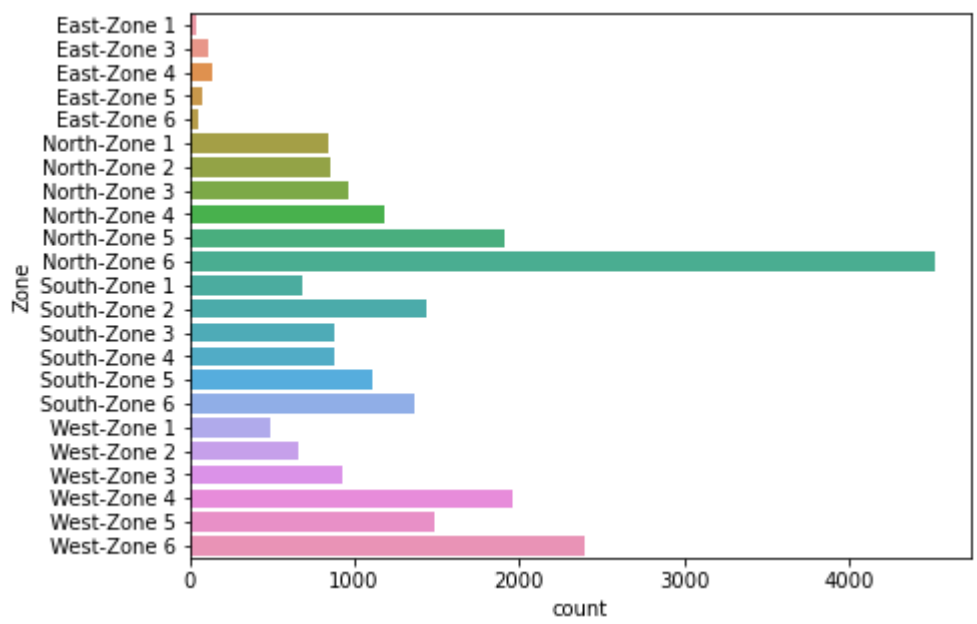
Categorical Features:

Countplot for variable: Location_type

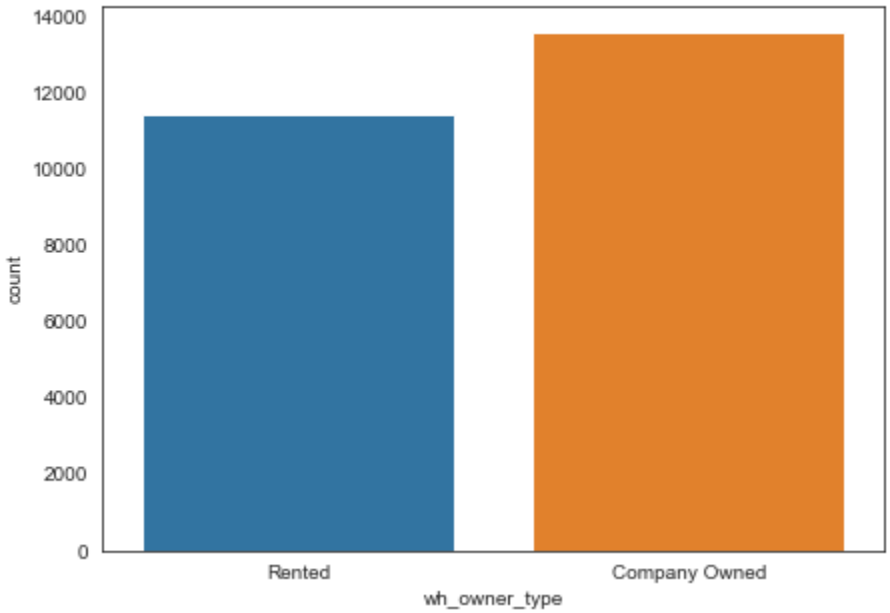


Countplot for variable: WH_capacity_size

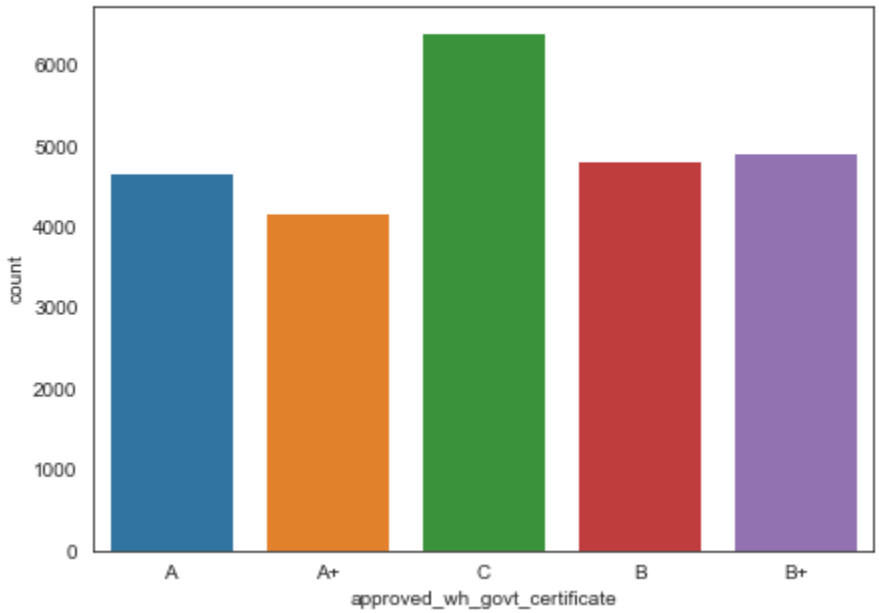




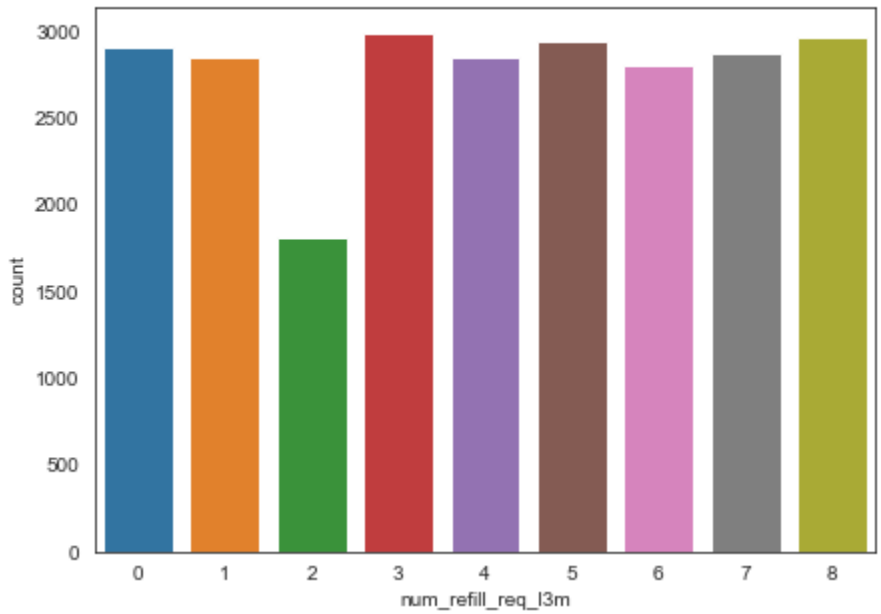
Countplot for variable: wh_owner_type



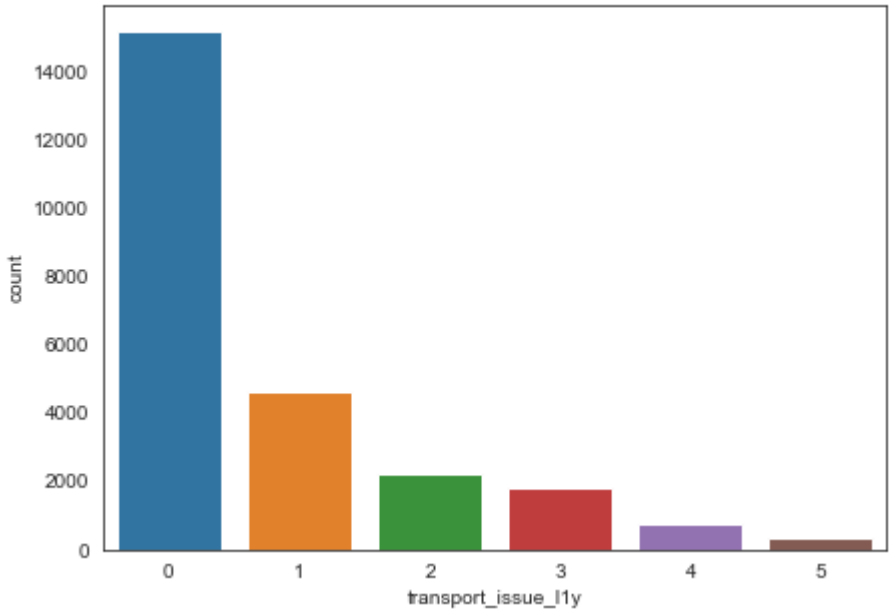
Countplot for variable: approved_wh_govt_certificate



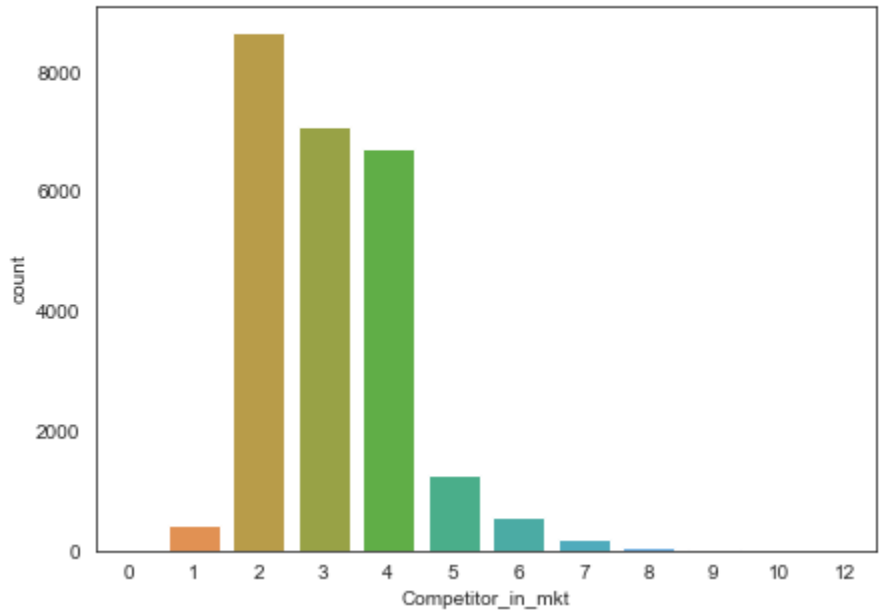
Countplot for variable: num_refill_req_l3m



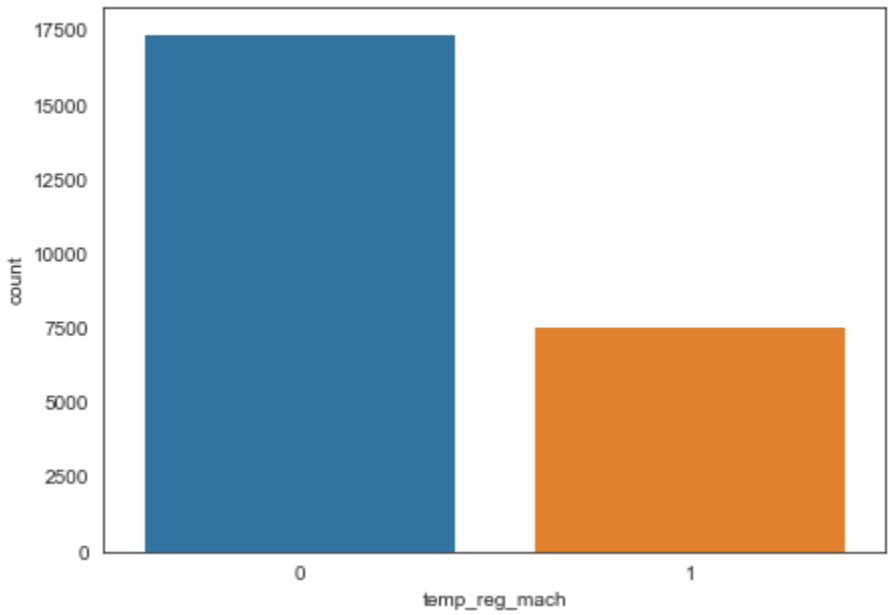
Countplot for variable: transport_issue_l1y



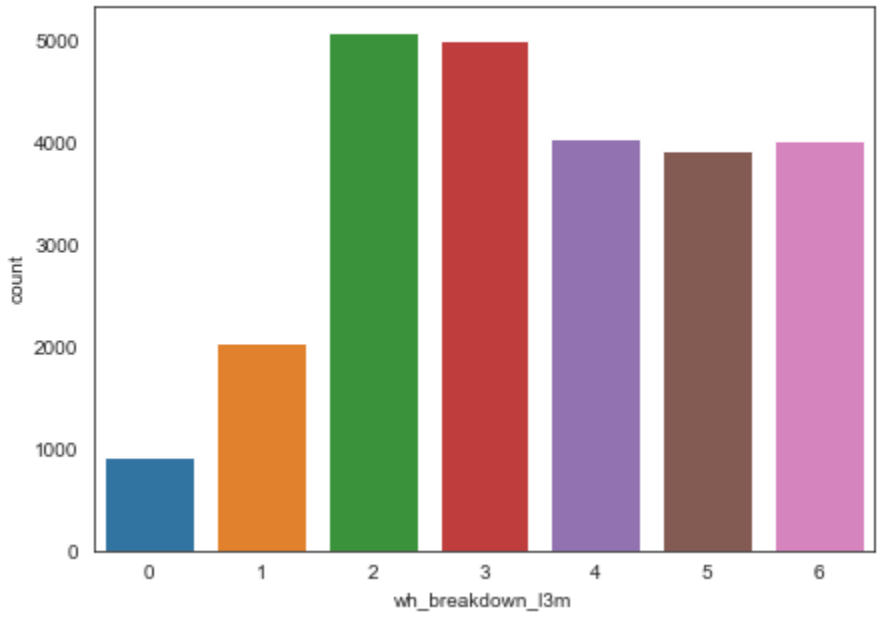
Countplot for variable: Competitor_in_mkt



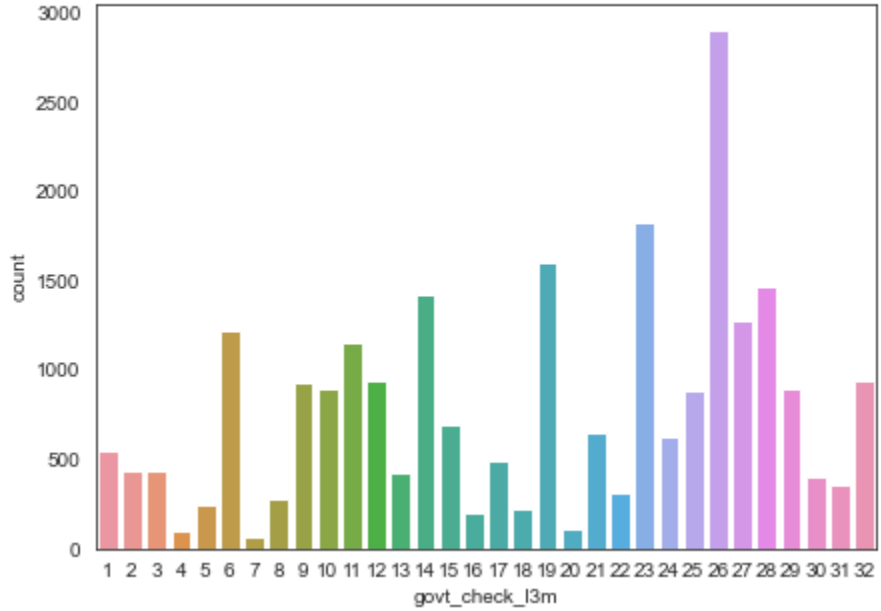
Countplot for variable: temp_reg_mach



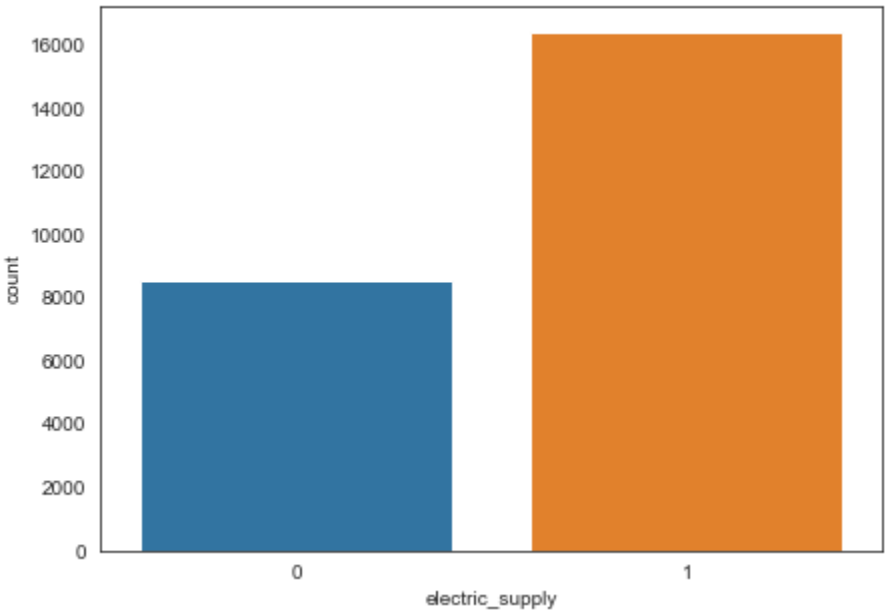
Countplot for variable: wh_breakdown_l3m



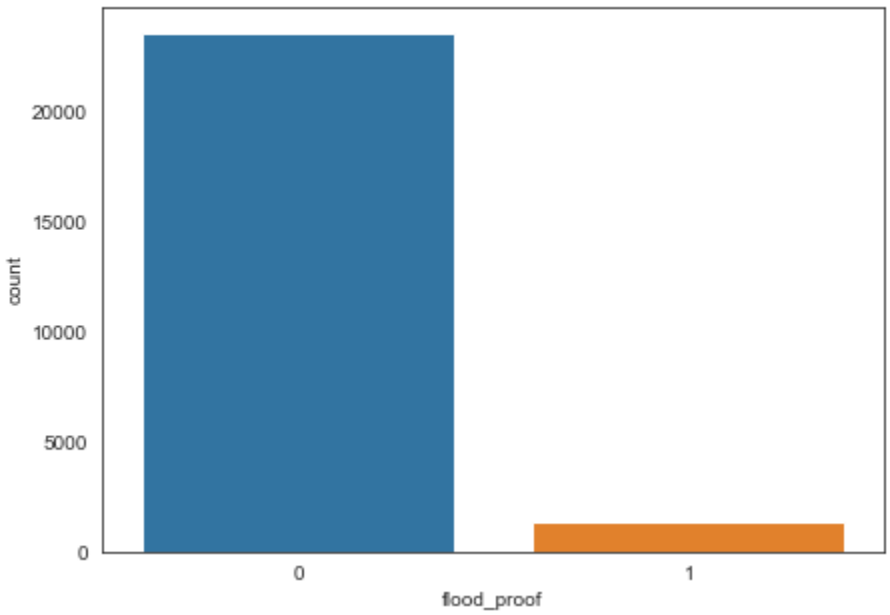
Countplot for variable: govt_check_l3m



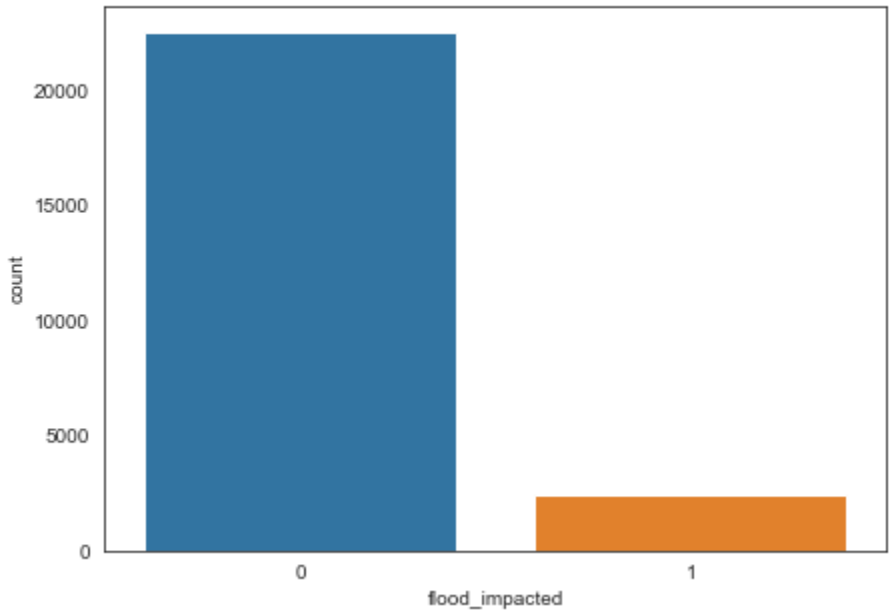
Countplot for variable: electric_supply



Countplot for variable: flood_proof



Countplot for variable: flood_impacted



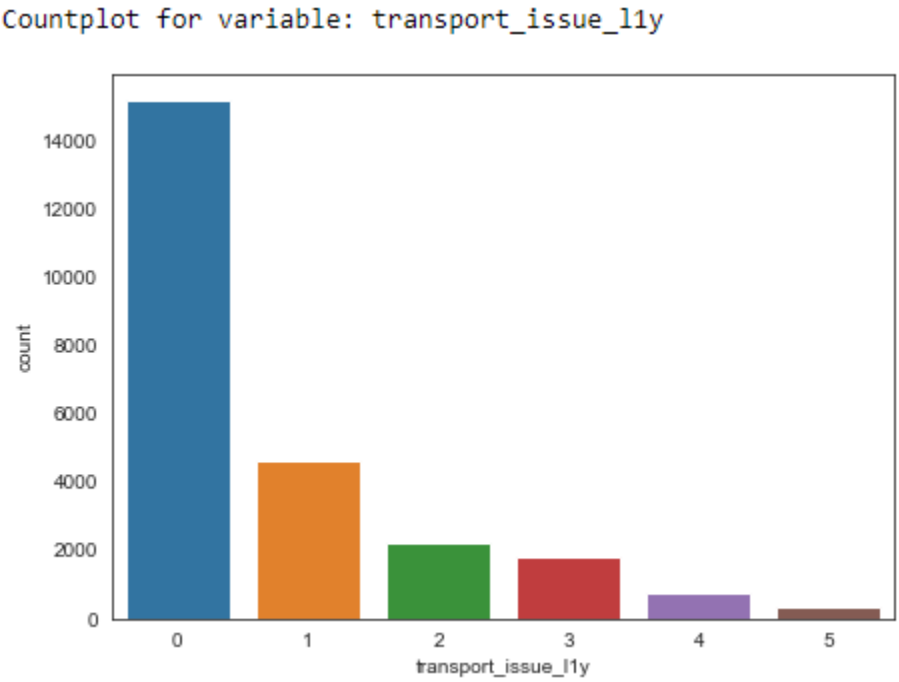


Figure 10: Countplots for categorical variables

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:

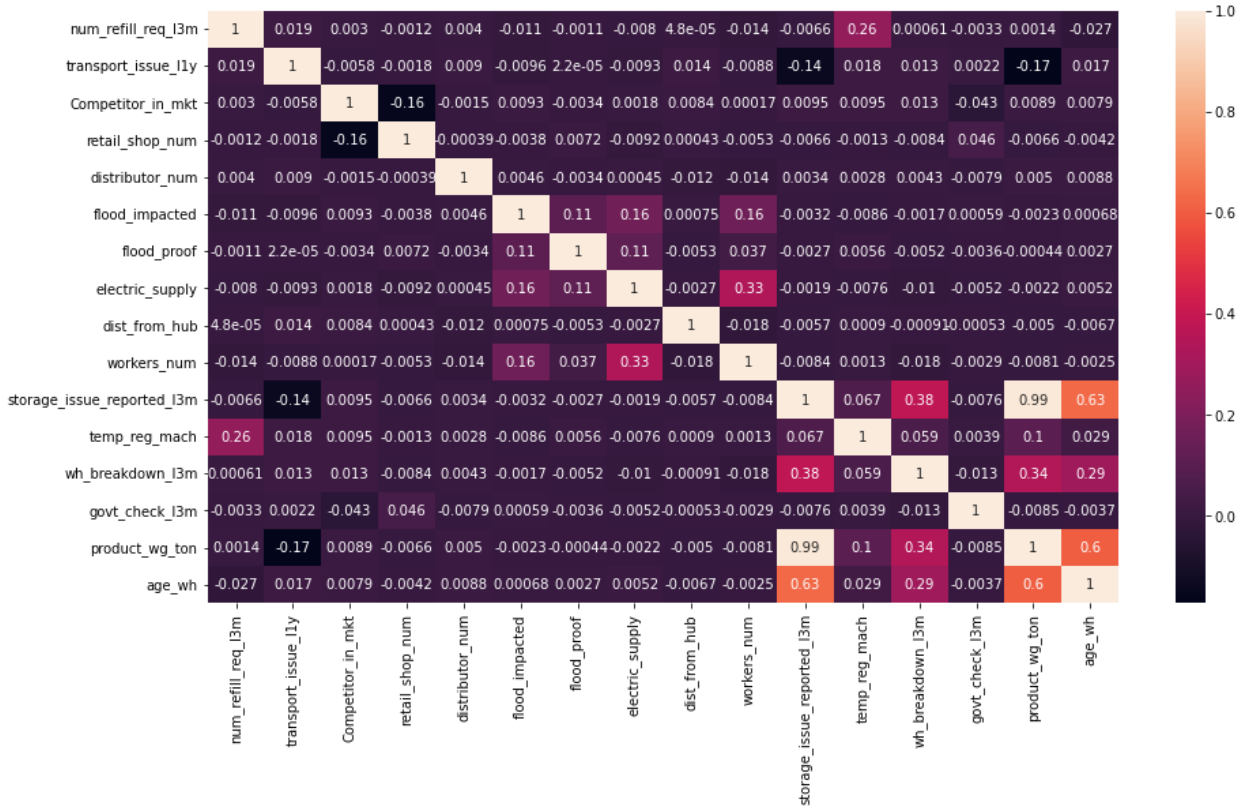


Figure 11: Correlation plot for numeric features present in the dataset

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:

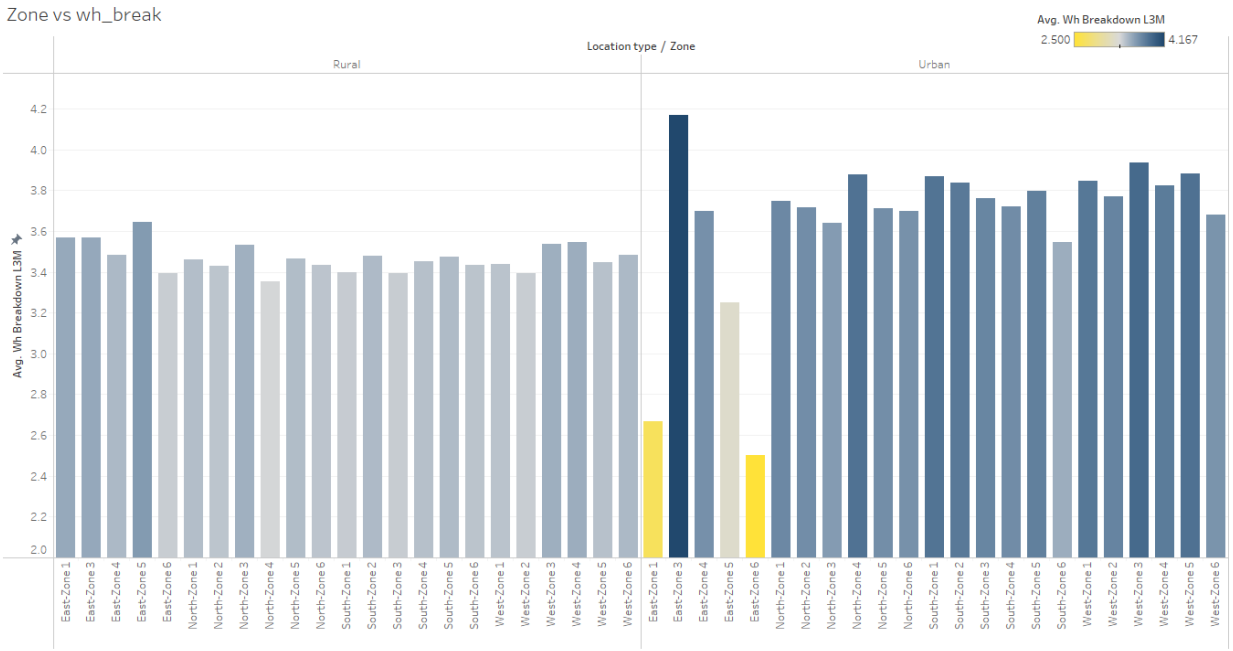


Figure 12: Zone vs. Warehouse Breakdown with location type

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.

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):

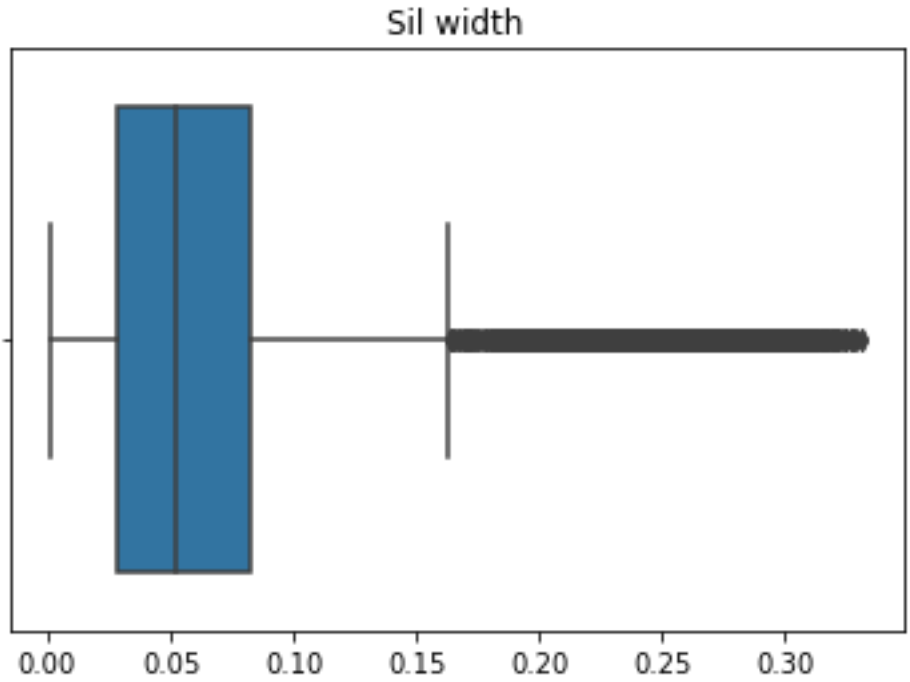


Figure 13: Silhouette width calculated based on identified clusters

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

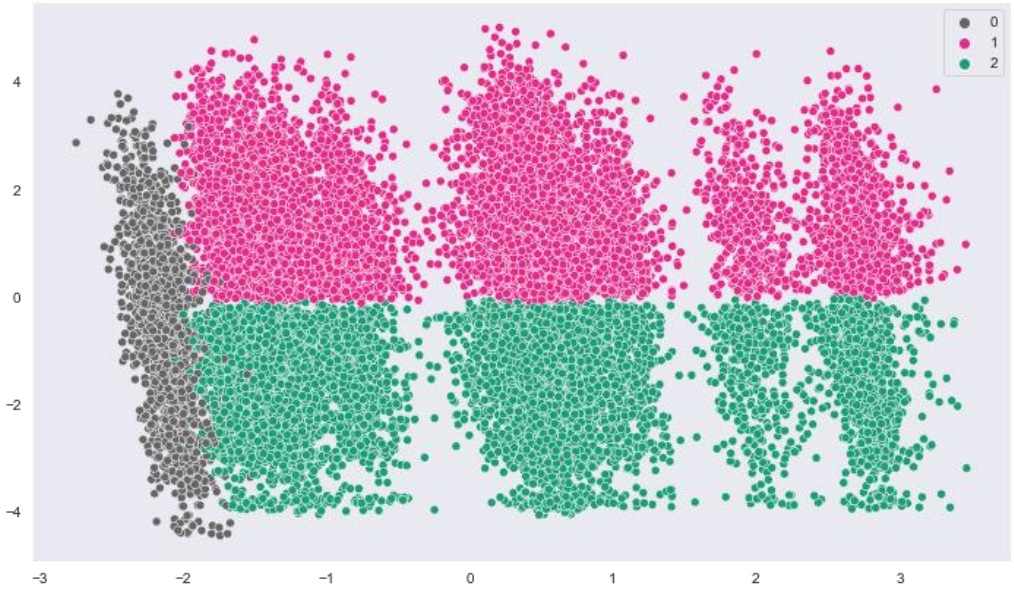


Figure 14: Visualization of formed clusters

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.