SUPPLY CHAIN PROJECT

NAME: VIKRAM REDDY KARLA

PGP-DSBA ONLINE

APRIL’22

DATE: 09-04-2023

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1) INTRODUCTION

1.a - Defining problem statement

A FMCG company which is two years old has been doing instant noodles business in several geographical locations. In an analysis it is found that there is a miss match in the demand and supply. Where the demand is high, supply is low and where the demand is low, supply is pretty high.

In both the cases it is an inventory cost loss to the company. hence, the higher management wants to optimize the supply quantity in each warehouse in entire country.

1.b - Need of the study/project

Due to variation in supply quantity, there is a case where an overloading of a Warehouse where there is low demand leading to oversupply. And vice versa where there is a case of underloading of supplies to warehouse leading undersupply.

From the current data we need to understand the Warehouses which have oversupplied & undersupplied and to build a model based on the data, which can optimize supply quantity to the warehouse.

1.c - Understanding business/social opportunity

Business understanding of the project is to increase the number of distributors between warehouse & the retail shops with a minimal effort, while keeping the product weight optimized which is being shipped to the warehouse.

The social opportunity is to acquire more clients and generating more revenue by keeping product weight optimized with the help of ML Models that are generated.

1.d - Understanding Data

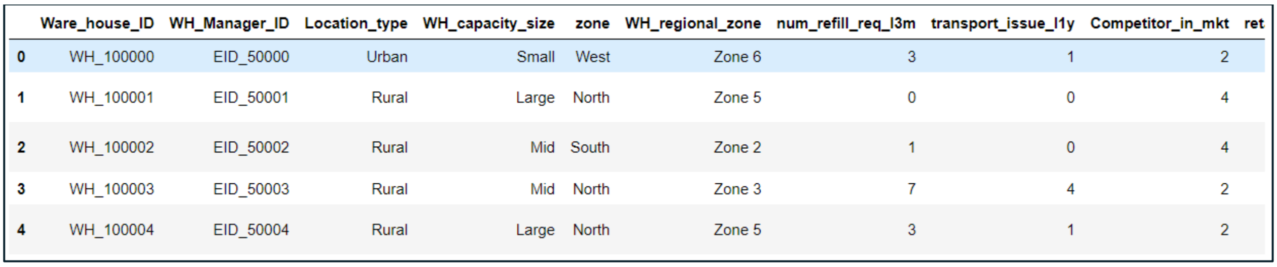
1.d.1 - Visual inspection of data (rows, columns, descriptive details)

Visual inspection of data found to be categorical & numerical data containing each individual Warehouse data.

The number of Rows in dataset : 25000

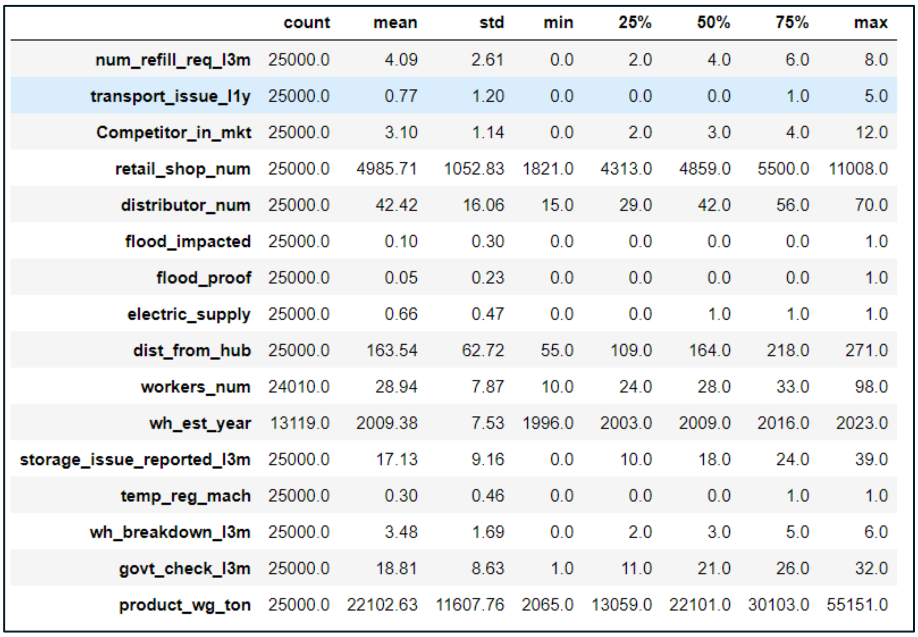
The number of Columns in dataset: 24

TABLE 1-1: SAMPLE DATASET

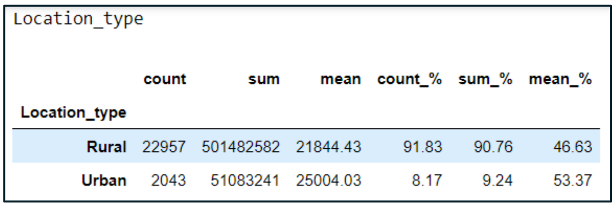


	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	ret
0	WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	2	
1	WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	4	
2	WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	4	
3	WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	2	
4	WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	2	

TABLE 1-2: DESCRIPTIVE STAISTICS

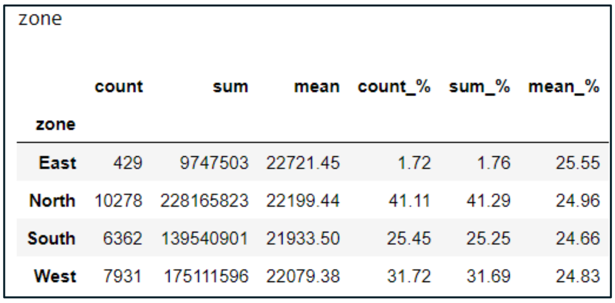


	count	mean	std	min	25%	50%	75%	max
num_refill_req_l3m	25000.0	4.09	2.61	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	0.77	1.20	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	3.10	1.14	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	4985.71	1052.83	1821.0	4313.0	4859.0	5500.0	11008.0
distributor_num	25000.0	42.42	16.06	15.0	29.0	42.0	56.0	70.0
flood_impacted	25000.0	0.10	0.30	0.0	0.0	0.0	0.0	1.0
flood_proof	25000.0	0.05	0.23	0.0	0.0	0.0	0.0	1.0
electric_supply	25000.0	0.66	0.47	0.0	0.0	1.0	1.0	1.0
dist_from_hub	25000.0	163.54	62.72	55.0	109.0	164.0	218.0	271.0
workers_num	24010.0	28.94	7.87	10.0	24.0	28.0	33.0	98.0
wh_est_year	13119.0	2009.38	7.53	1996.0	2003.0	2009.0	2016.0	2023.0
storage_issue_reported_l3m	25000.0	17.13	9.16	0.0	10.0	18.0	24.0	39.0
temp_reg_mach	25000.0	0.30	0.46	0.0	0.0	0.0	1.0	1.0
wh_breakdown_l3m	25000.0	3.48	1.69	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	18.81	8.63	1.0	11.0	21.0	26.0	32.0
product_wg_ton	25000.0	22102.63	11607.76	2065.0	13059.0	22101.0	30103.0	55151.0



Location_type

	count	sum	mean	count_%	sum_%	mean_%
Location_type						
Rural	22957	501482582	21844.43	91.83	90.76	46.63
Urban	2043	51083241	25004.03	8.17	9.24	53.37



zone

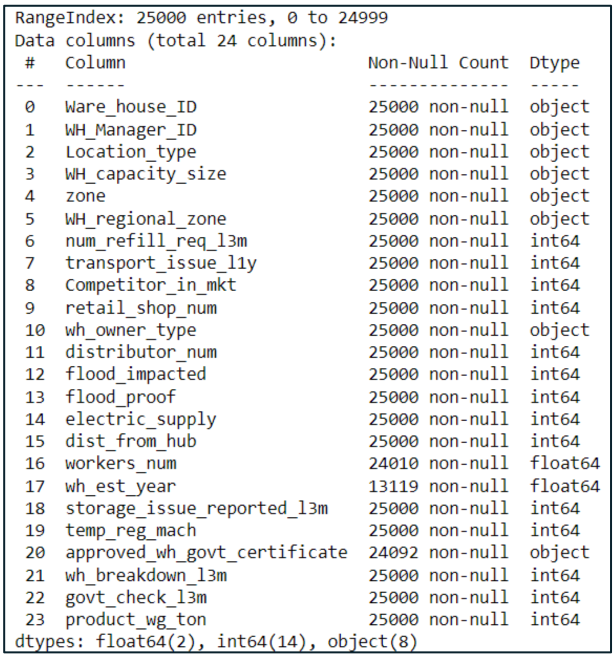
	count	sum	mean	count_%	sum_%	mean_%
zone						
East	429	9747503	22721.45	1.72	1.76	25.55
North	10278	228165823	22199.44	41.11	41.29	24.96
South	6362	139540901	21933.50	25.45	25.25	24.66
West	7931	175111596	22079.38	31.72	31.69	24.83

- Electricity supply, flood impacted, flood proof, temperature machine is either 0 or 1, which means ordinal data.
- Storage issue reported, warehouse breakdown, government checks, transport issues, refills, competitors etc are found to be discrete data.

- Number of workers, number of distributors, distance from production to warehouse are between 0 to 300. And considering it as discrete data because of the uniqueness in data.
- Product weight in tons is a continuous data with minimum = 2065, mean = 22102 & max= 55151.
- More than 90% of product weight is being shipped & distributed to Rural areas.
- West, North & South are the regions where more than 90% of product weight is shipped & distributed.

1.d.2 - Understanding of attributes (variable info, renaming if required)

TABLE 1-3: DATA INFO



```
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)
```

- Data consists of integer, floating and object data.
- A year attribute is observed.
- No duplicates were observed.
- Null values are observed in workers_num, warehouse year, govt_certificate.
- Column name looks fine renaming not required.

2 - EXPLORATORY DATA ANALYSIS

2.a - Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

Most of the attributes are Nominal, Ordinal & discrete data. Hence univariate analysis is performed for Product weight which is found to be continuous data. The univariate analysis is also done for distance from hub & Number of workers.

Description of workers_num

- count 24010.000
- mean 28.944
- std 7.873
- min 10.000
- 25% 24.000
- 50% 28.000
- 75% 33.000
- max 98.000
- Skewness 1.06
- Kurtosis 3.409
- W = nan p_value = 1.0 workers_num P_value > 0.05 is Normal

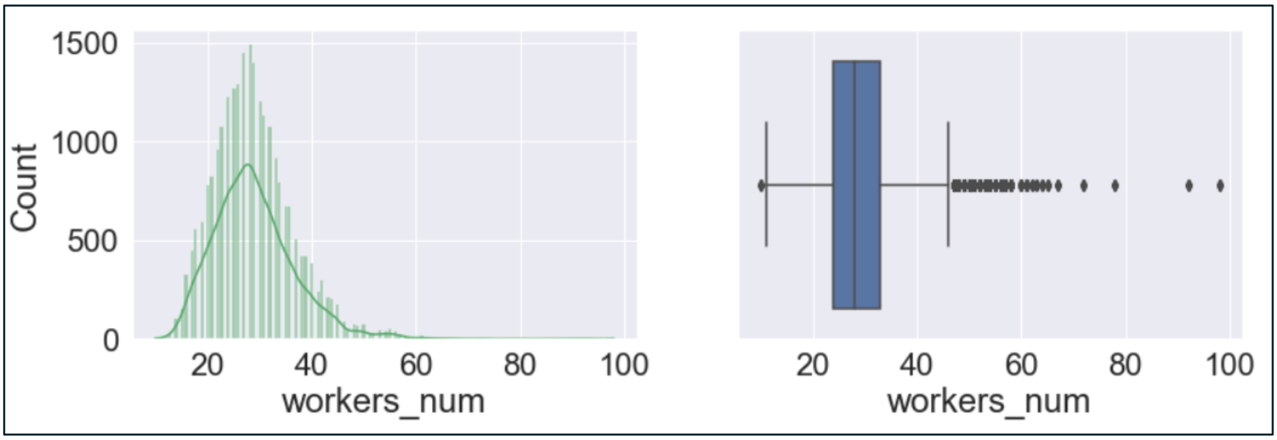


FIGURE 2-1: NUMNBER OF WORKERS DISTRIBUTION PLOT

- It is normally distributed.
- Outliers' presence is very much and on right whisker.

Description of dist_from_hub

- ---
- count   25000.000
- mean   163.537
- std   62.719
- min   55.000
- 25%   109.000
- 50%   164.000
- 75%   218.000
- max   271.000
- Skewness -0.006
- Kurtosis -1.201
- W = 0.955 p_value = 0.0 dist_from_hub P_value < 0.05 is NOT Normal

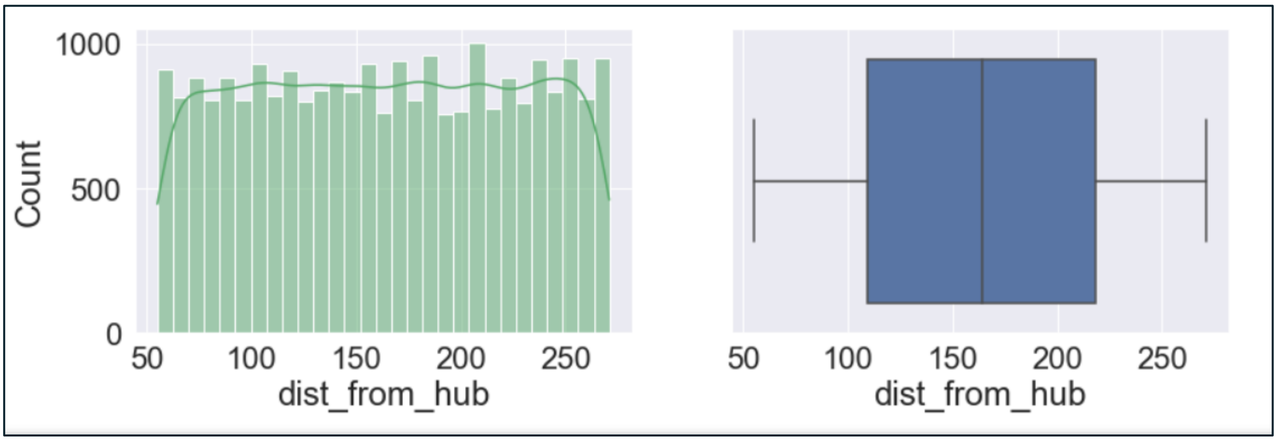


FIGURE 2-2: DISTANCE FROM HUB DISTRIBUTION PLOT

- Rectangular distribution, not useful.

Description of product_wg_ton

- count   25000.000
- mean   22102.633
- std   11607.755
- min   2065.000
- 25%   13059.000
- 50%   22101.000
- 75%   30103.000
- max   55151.000
- Skewness 0.332

- Kurtosis -0.502
- W = 0.971 p_value = 0.0 product_wg_ton P_value < 0.05 is NOT Normal

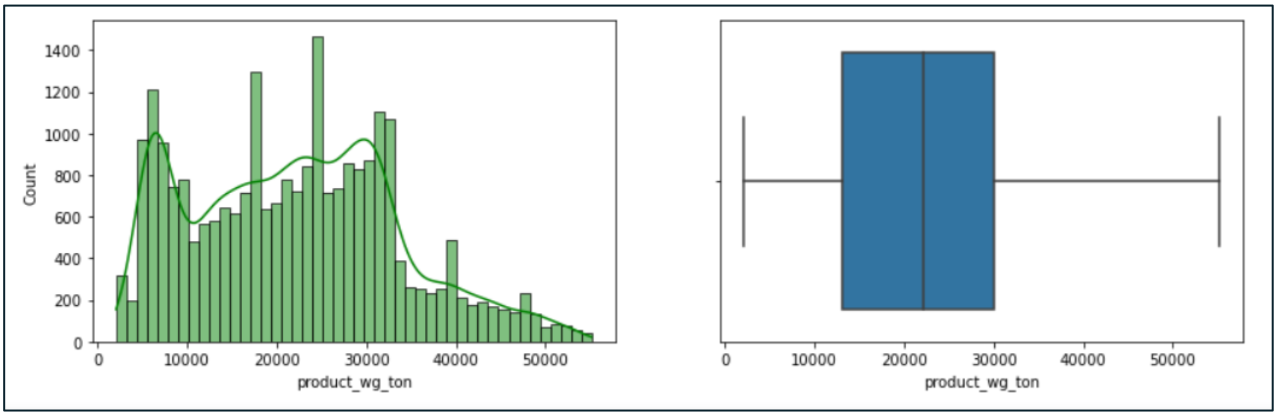


FIGURE 2-3: PRODUCT WEIGHT DISTRIBUTION PLOT

- It is observed to be bit bi modal & right skewed data for the product weight.
- minimum = 2065, mean = 22102 & max= 55151.
- the data is not normally distributed.
- No outliers were present.
- No null values were present.

2.b - Bivariate analysis (relationship between different variables, correlations)

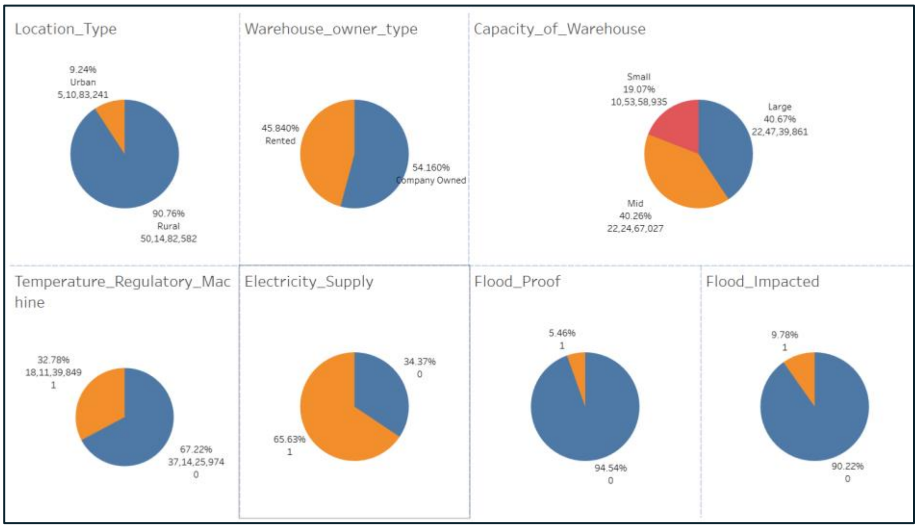


FIGURE 2-4: BINARY & TRINARY DATA PIE PLOT

- More than 90% product weight shipped to warehouse have been designed for flood proof & flood impacted is 0.
- More than 65% supply weights does not have temperature regulatory machines & have additional electrical supply.

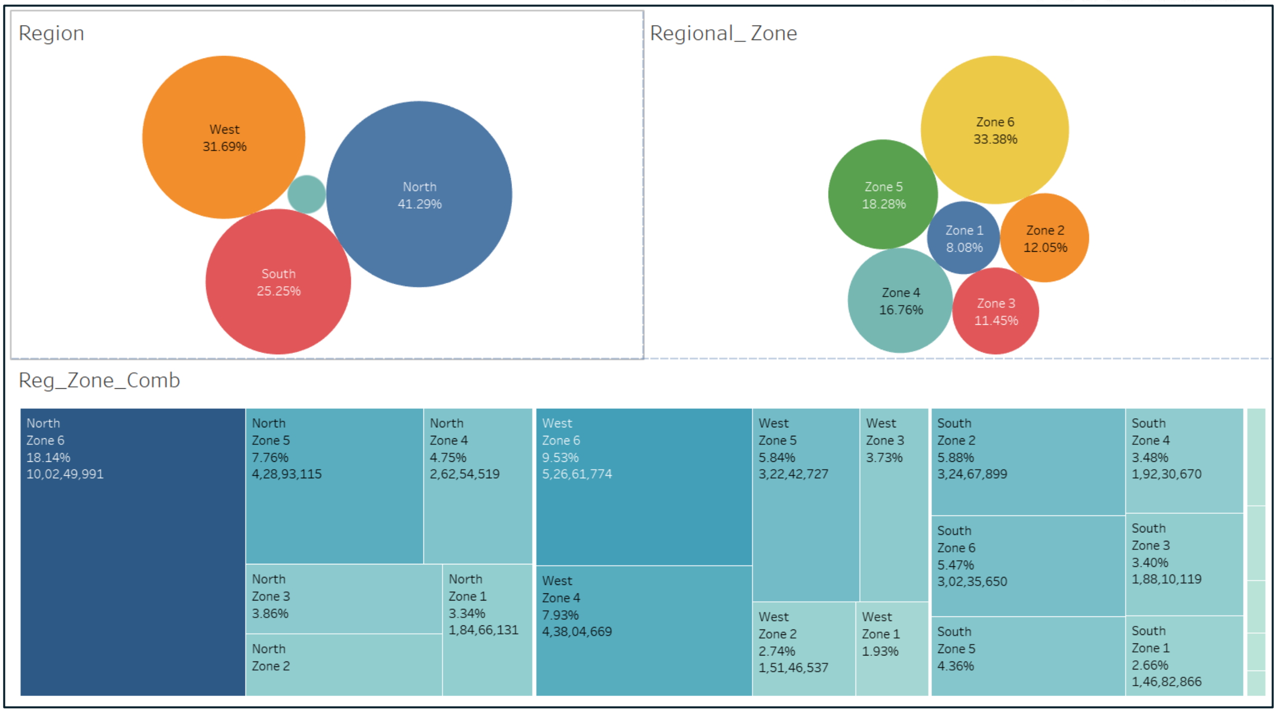
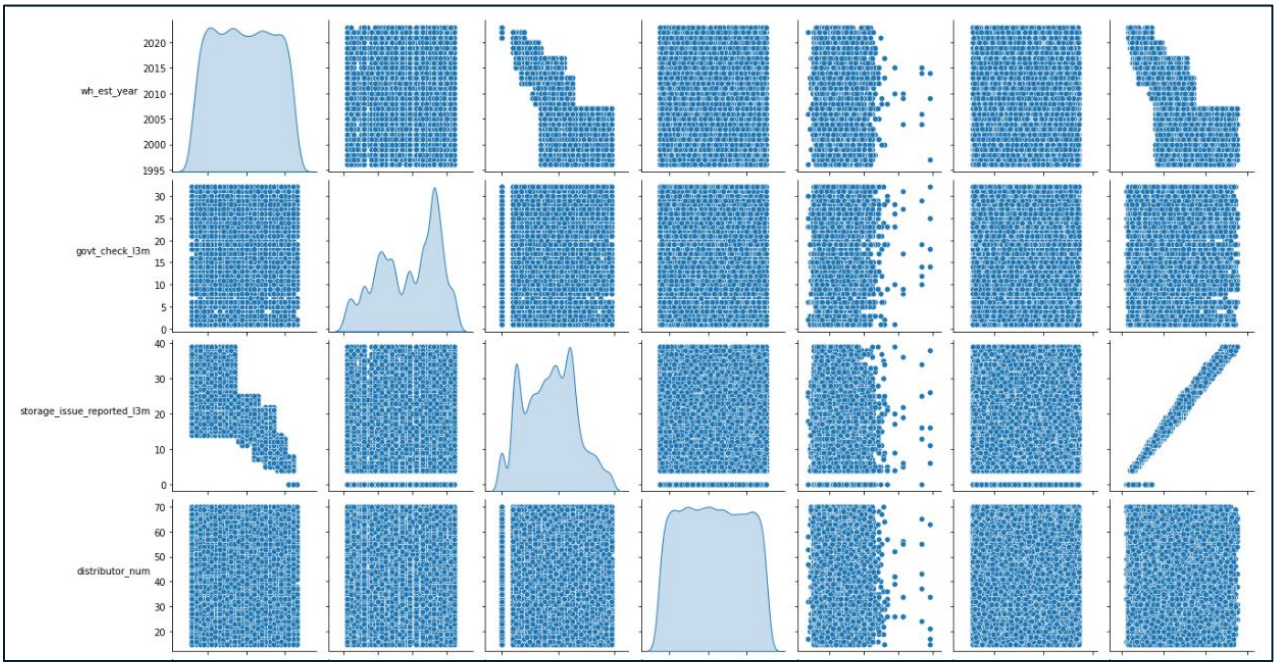


FIGURE 2-5: REGION & REGIONAL ZONES PLOT

- Zone 6 of North region alone constitutes for 18% of product weight shipping to warehouses.



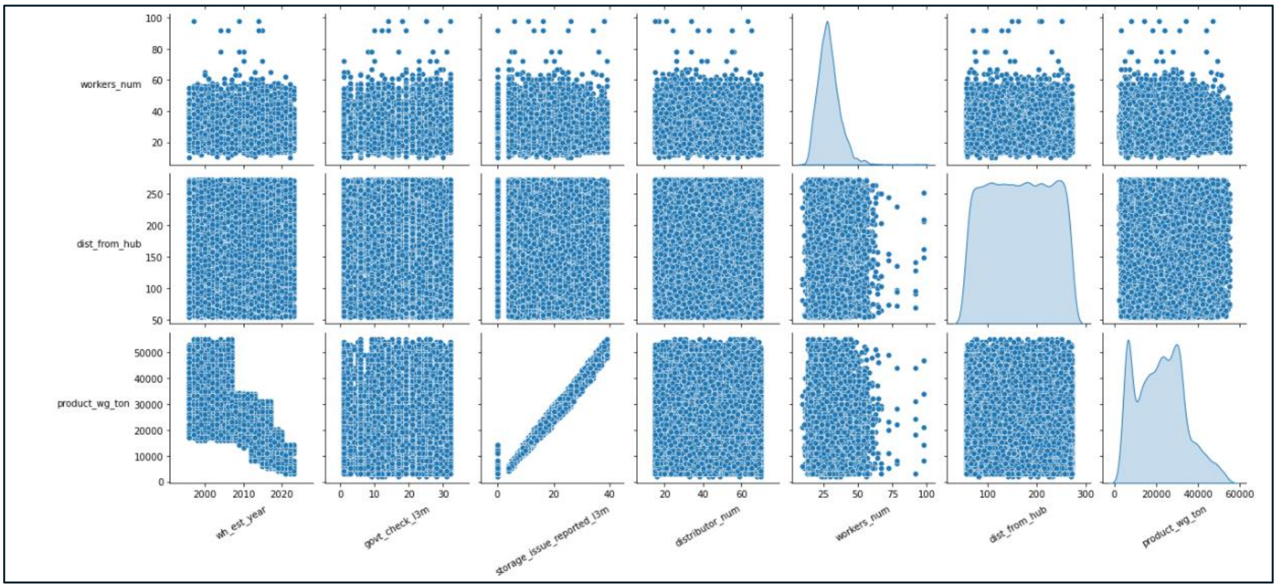
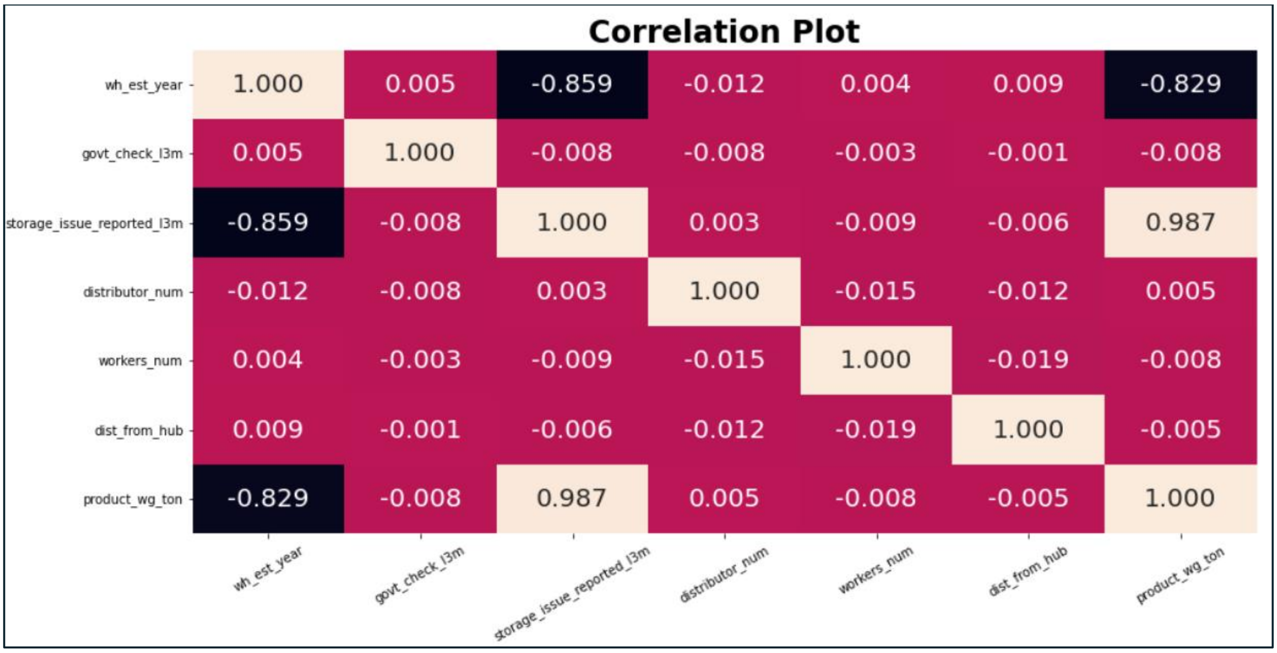


FIGURE 2-6: BIVARIATE ANALYSIS PLOT



Correlation Plot

	wh_est_year	govt_check_l3m	storage_issue_reported_l3m	distributor_num	workers_num	dist_from_hub	product_wg_ton
wh_est_year	1.000	0.005	-0.859	-0.012	0.004	0.009	-0.829
govt_check_l3m	0.005	1.000	-0.008	-0.008	-0.003	-0.001	-0.008
storage_issue_reported_l3m	-0.859	-0.008	1.000	0.003	-0.009	-0.006	0.987
distributor_num	-0.012	-0.008	0.003	1.000	-0.015	-0.012	0.005
workers_num	0.004	-0.003	-0.009	-0.015	1.000	-0.019	-0.008
dist_from_hub	0.009	-0.001	-0.006	-0.012	-0.019	1.000	-0.005
product_wg_ton	-0.829	-0.008	0.987	0.005	-0.008	-0.005	1.000

FIGURE 2-7: CORRELATION PLOT

- There is a strong positive relation between storage issue reported & product weight.
- There is strong negative relation between warehouse year & product weight.
- There is strong negative relation between warehouse year & storage issue reported.

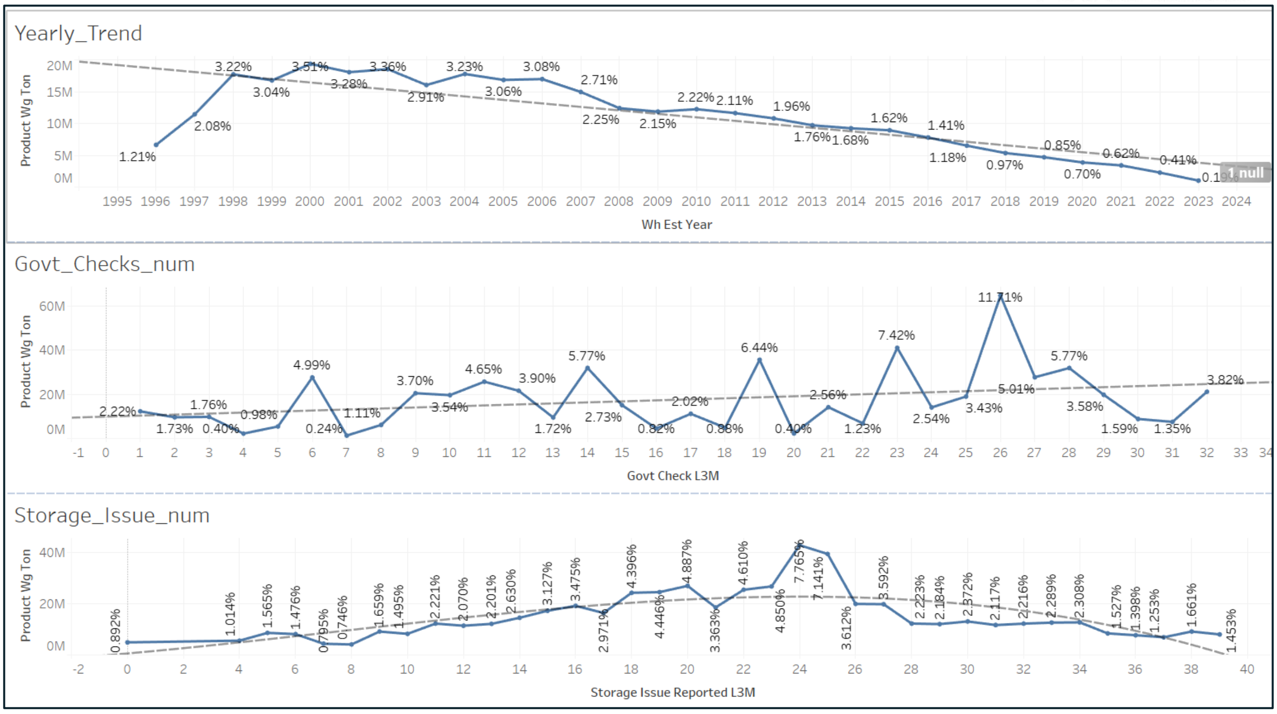


FIGURE 2-8: DISCRETE DATA PRODUCT WEIGHT TREND PLOT

- By keeping Storage issue, govt checks & established year as discrete data and plotting trend with product weight.
- There is downtrend with established warehouse year.
- Strong uptrend if number of govt checks increases.
- Some kind of poly trend with storage issue reported.

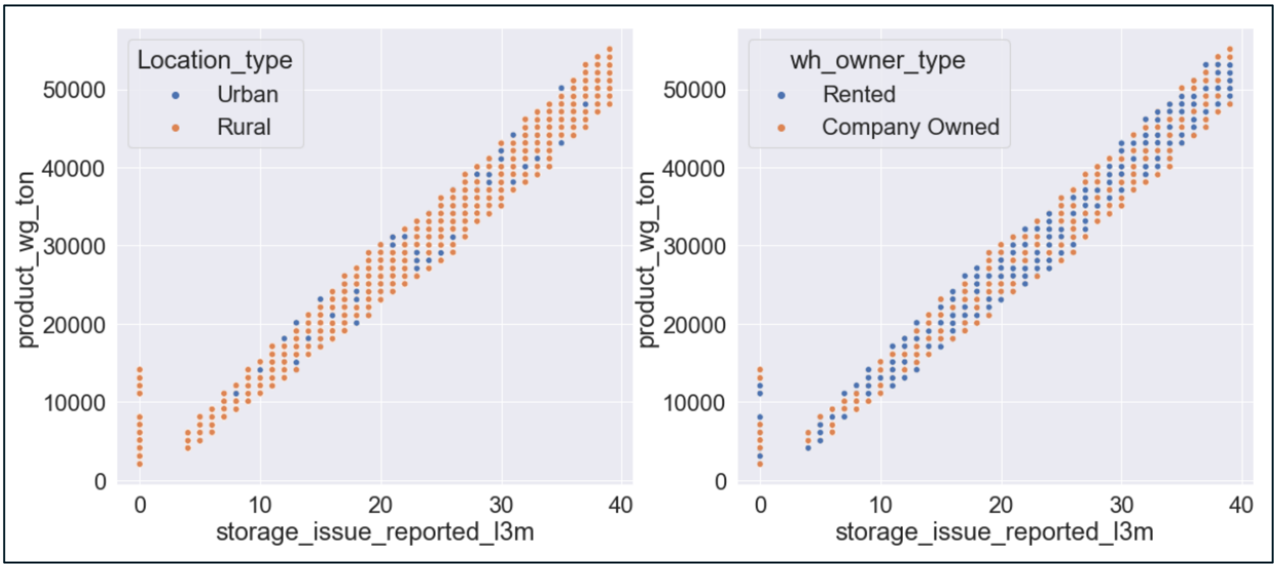


FIGURE 2-9: MULTIVARIATE PLOT

2.c - Business insights from EDA

2.c.1 - business insights using clustering.

Clustering is performed only on original dataset without the new variables.

Clustering has performed and 3 cluster were chosen by using silhouette_score

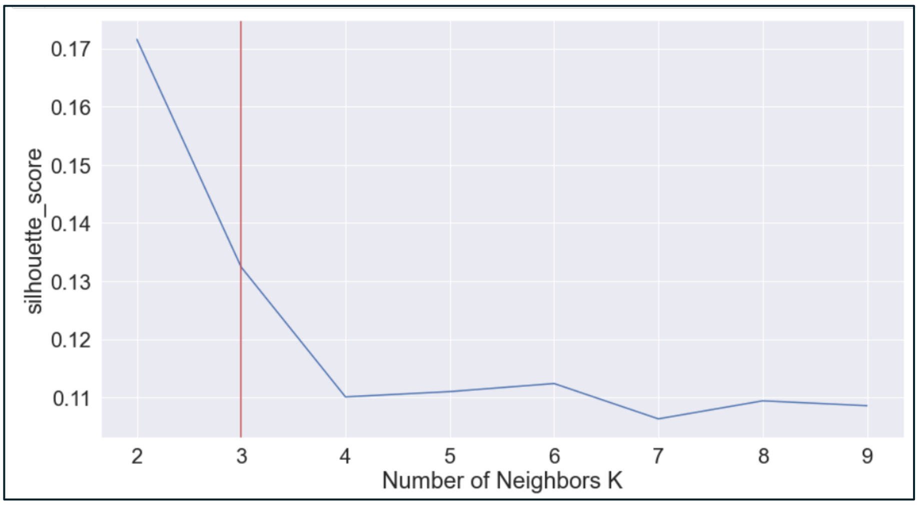
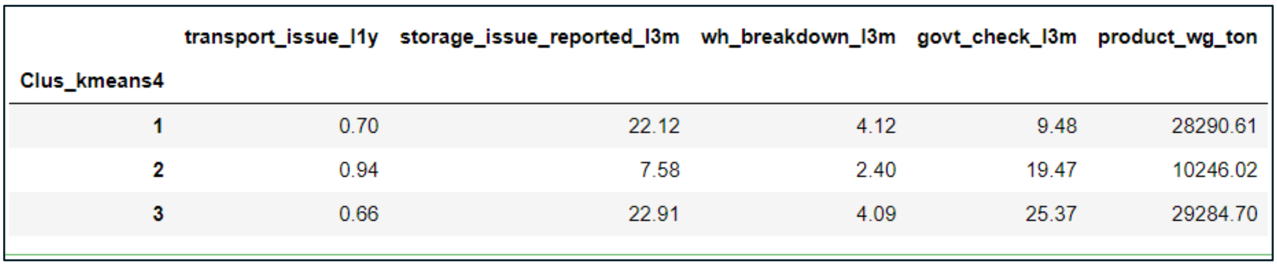


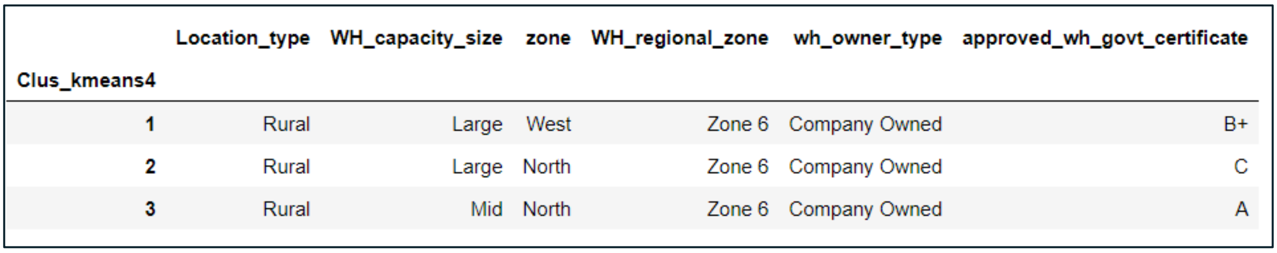
FIGURE 2-10: SILHOUTTE SCORE PLOT

TABLE 2-1: NUMERICAL VARIATION CLUSTER



	transport_issue_l1y	storage_issue_reported_l3m	wh_breakdown_l3m	govt_check_l3m	product_wg_ton
Clus_kmeans4					
1	0.70	22.12	4.12	9.48	28290.61
2	0.94	7.58	2.40	19.47	10246.02
3	0.66	22.91	4.09	25.37	29284.70

TABLE 2-2: CATEGORICAL VARIATION CLUSTER



	Location_type	WH_capacity_size	zone	WH_regional_zone	wh_owner_type	approved_wh_govt_certificate
Clus_kmeans4						
1	Rural	Large	West	Zone 6	Company Owned	B+
2	Rural	Large	North	Zone 6	Company Owned	C
3	Rural	Mid	North	Zone 6	Company Owned	A

- 1 & 3 cluster are equal in product weight, but cluster 2 is less than half of any other two clusters.
- Government checks are high for cluster 3 >2>1.
- Cluster 2 has lowest storage issues & higher transport issues.
- Cluster 1 & 2 are large sized.
- Cluster 3 has highest approved govt certificate.

2.c.2 - Business Insights from EDA

- 90% of product weight is shipped to Rural area, so marketing & Ads can be increase in Rural areas.
- North & West in combination with Zone 5/6 caters for more than 50% of product weight. Thes reginal zones can be looked for more sale's intensive programs.

- Storage Issue reported in last three months has a strong correlation with product weight, which can be used to determine product weight to be shipped.
- Large & Mid warehouses account for more than 80% of weight, hence number of warehouses can be increased.

2.c.3 - Other business insights

- It’s an instant noodle business, very customer centric and customer should be top priority.
- Assure safety, security, hygiene of the warehouses.
- Urban areas can have higher average, but its rural areas are most amount of produce to be shipped, therefore increase number of Warehouses strategically & geolocations where towns, inter cities were present.
- Cost effective transportation like electric vehicles can used to reduce operation costs, for example smart trucks.
- Local employees can be trained for warehouse or transportation works.
- Strong brand image & brand equity.
- Good advertising & visibility.

3 - DATA CLEANING AND PRE-PROCESSING

3.a - Approach

- Few attributes were not required, hence finding them & dropping them.
- Missing values are imputed with KNN Imputer.
- Columns with missing values more than 40% are dropped.
- The continuous variables which are distance from hub & Product weight do not have outliers. So, no outlier treatment.
- All categorical attributes are label encoded.

3.b - Missing Value treatment

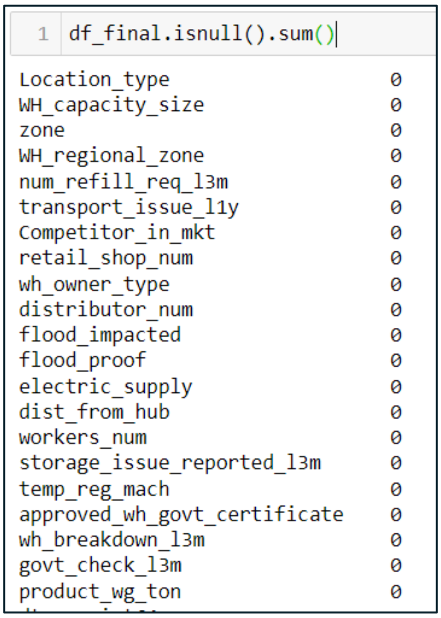
There are missing values present in

- wh_est_year
- workers_num
- approved_wh_govt_certificate

wh_est_year has been dropped due to 47% missing values; imputing creates huge bias.

workers_num, approved_wh_govt_certificate attributes missing values are treated with KNN imputer.

TABLE 3-1: TABLE AFTER NULL VALUE IMPUTATION



```
df_final.isnull().sum()
```

```
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                     0
storage_issue_reported_l3m      0
temp_reg_mach                   0
approved_wh_govt_certificate    0
wh_breakdown_l3m                0
govt_check_l3m                  0
product_wg_ton                  0
```

3.c - Outlier treatment

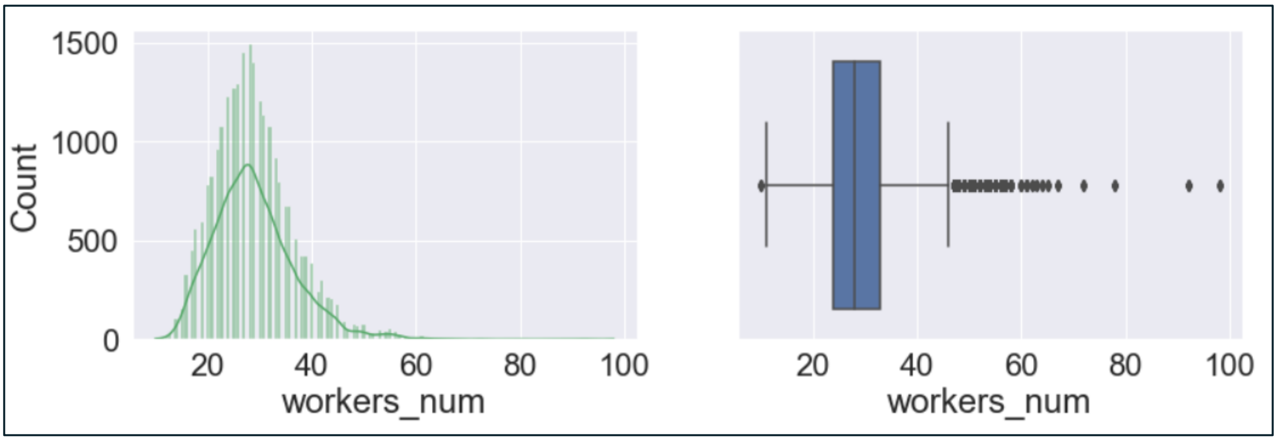


FIGURE 3-1: NUMNBER OF WORKERS DISTRIBUTION PLOT

- Most importantly continuous variables like Distance from hub & Product weight do not have outliers. Hence No outlier treatment. **Refer *Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)***

- From number of workers distribution plot, it is clear that this almost continuous data has outliers, In perspective business NO bias to be introduced at the beginning stage. Most importantly Number of workers is discrete data.
- So, therefore No outlier treatment.

3.d - Need for variable transformation

Label Encoding has been performed on all categorical variables which are object. The variable transformation is performed for below variables.

'Location_type',

'WH_capacity_size',

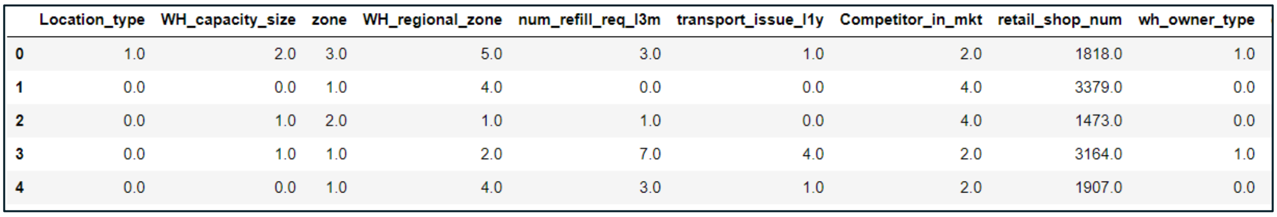
'zone',

'WH_regional_zone',

'wh_owner_type',

'approved_wh_govt_certificate'

TABLE 3-2: LABEL ENCODED TABLE

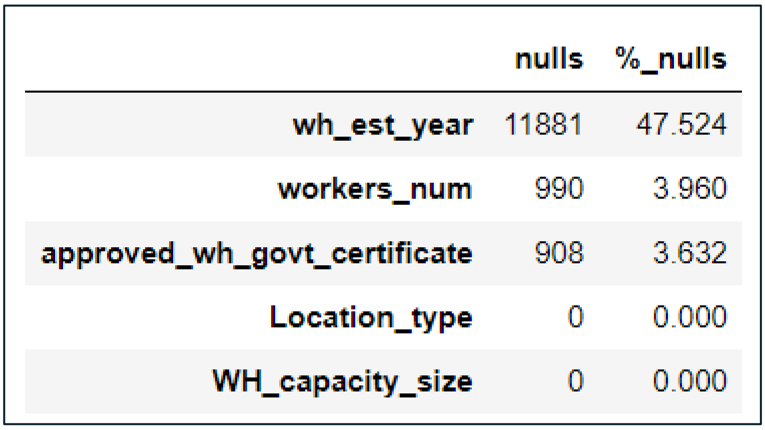


	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	wh_owner_type
0	1.0	2.0	3.0	5.0	3.0	1.0	2.0	1818.0	1.0
1	0.0	0.0	1.0	4.0	0.0	0.0	4.0	3379.0	0.0
2	0.0	1.0	2.0	1.0	1.0	0.0	4.0	1473.0	0.0
3	0.0	1.0	1.0	2.0	7.0	4.0	2.0	3164.0	1.0
4	0.0	0.0	1.0	4.0	3.0	1.0	2.0	1907.0	0.0

3.e - Variables removed or added and why (if any)

Ware_house_ID, WH_Manager_ID are the unwanted variables which are not required.

TABLE 3-3: NULL VALUE TABLE



	nulls	%_nulls
wh_est_year	11881	47.524
workers_num	990	3.960
approved_wh_govt_certificate	908	3.632
Location_type	0	0.000
WH_capacity_size	0	0.000

- More than 47% of null values are present in warehouse establishment year, which on imputation is very highly biased and hence putting this attribute in unwanted list.

Ware_house_ID, WH_Manager_ID, warehouse establishment year are the unwanted variables which are not required and hence dropping from the data frame.

4 - MODEL BUILDING AND INTERPRETATION.

4.a - Building various models (approach)

From the dataset the Target variable is "***product_wg_ton***". Which is a continuous feature.

Hence, a **Regression model** is **to be built** to optimize the weight of the product that is being sent to the Warehouse.

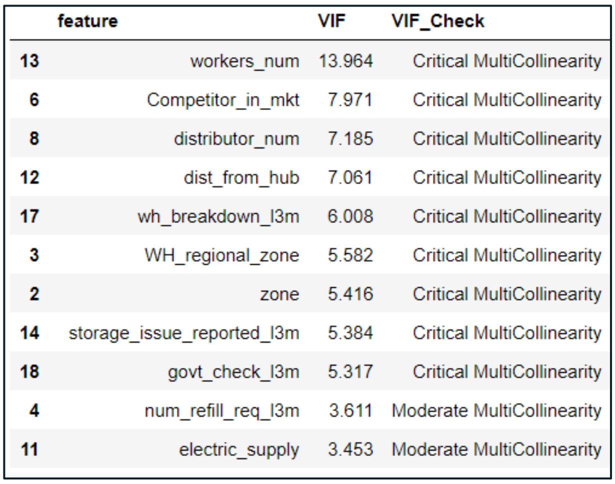
4.a.1 - Types of models (chosen models)

There are several different algorithms can be used to fit & predict the regression problem. In this case it has been chosen in four different varieties and each algorithm from those varieties.

1. Generalized Linear Models → Linear Regression
2. Ensemble Model → Bagging type – RandomForestRegressor
3. Ensemble Model → Boosting type – XGBRegressor
4. Neural Network Model → MLPRegressor

4.a.2 - Multicollinearity check

TABLE 4-1: MULTICOLLINEARITY TABLE



	feature	VIF	VIF_Check
13	workers_num	13.964	Critical MultiCollinearity
6	Competitor_in_mkt	7.971	Critical MultiCollinearity
8	distributor_num	7.185	Critical MultiCollinearity
12	dist_from_hub	7.061	Critical MultiCollinearity
17	wh_breakdown_l3m	6.008	Critical MultiCollinearity
3	WH_regional_zone	5.582	Critical MultiCollinearity
2	zone	5.416	Critical MultiCollinearity
14	storage_issue_reported_l3m	5.384	Critical MultiCollinearity
18	govt_check_l3m	5.317	Critical MultiCollinearity
4	num_refill_req_l3m	3.611	Moderate MultiCollinearity
11	electric_supply	3.453	Moderate MultiCollinearity

From the above table it is clear that there is good amount of multicollinearity in the data. Hence Bagging (RandomForestRegressor) models were included to counter the multicollinearity and Boosting (XGBRegressor) models to counter overfitting.

The neural network can be helpful in case of strong correlation.

4.b - Model building

4.b.1 - Aproach

- **No New columns** were included.
- The train test split done with **70:30** ratio of **Train & Test**.
- **Scaling** is done only for **Neural Network model.**
- **GridsearchCV** was used to explore more with **Hyper Parameter Tuning** & **cross validation.**
- Model explain ability done with ***Shapley's Explanations*** to get the most important features.
- **Other** model explanations were also used like ***Sequential Feature Selection*** & ***Coefficients*** & ***Feature Importance's*** to check for the most important features.
- **Evaluation** metrics like **R-Squared, RMSE, MAPE** were used to test the Test dataset.

4.b.2 - Linear Regression

Linear Regression fitted on Train dataset and the results are shown below.

R-Squared (1- SSE/SST) Train data: 0.977

R-Squared (1- SSE/SST) Test data: 0.978

RMSE Train data: 1748.132

RMSE Test data: 1692.693

Mean Abs % Error Train data: 8.8 %

Mean Abs % Error Test data: 8.688 %

Feature Importance's

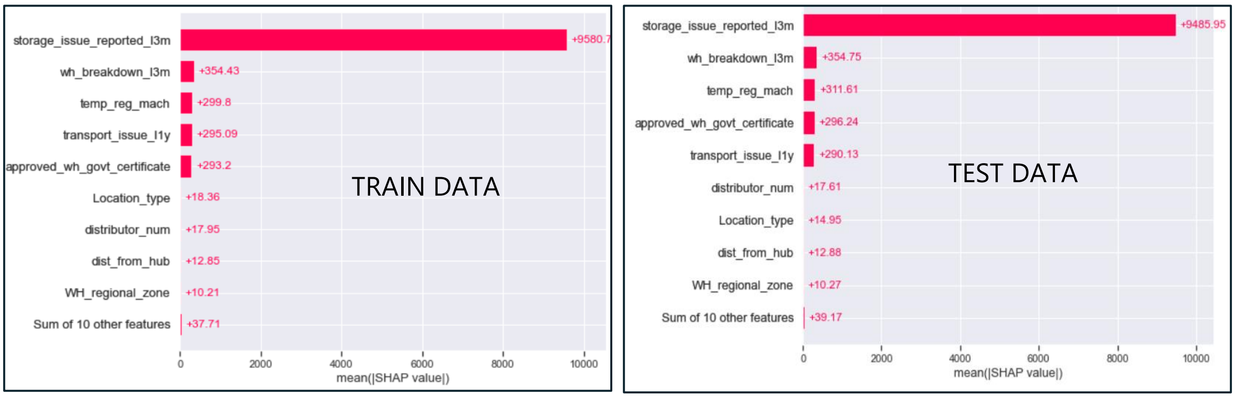


FIGURE 4-1: LINEAR REGRESSION FEATURE IMPORTANCE PLOT FOR TRAIN & TEST

Regression equation

Location type * -105.89 + WH capacity size * 8.842 + zone * -2.508 + WH regional zone * -7.125 + num refill req l3m * 4.342 + transport issue l1y * -312.236 + Competitor in mkt * -5.243 + wh owner type * 10.074 + distributor num * 1.272 + flood impacted * 8.325 + flood proof * 48.796 + electric supply * 8.443 + dist from hub * 0.235 + workers num * 0.152 + storage issue reported l3m * 1253.259 + temp reg mach * 736.869 + approved wh govt certificate * 246.818 + wh breakdown l3m * -243.695 + govt check l3m * -0.002

4.b.3 - RandomForestRegressor

RandomForestRegressor fitted on Train dataset and the results are shown below.

R-Squared (1- SSE/SST) Train data : 0.995

R-Squared (1- SSE/SST) Test data : 0.994

RMSE Train data : 852.301

RMSE Test data : 893.355

Mean Abs % Error Train data : 3.995 %

Mean Abs % Error Test data : 4.22 %

Feature Importance's

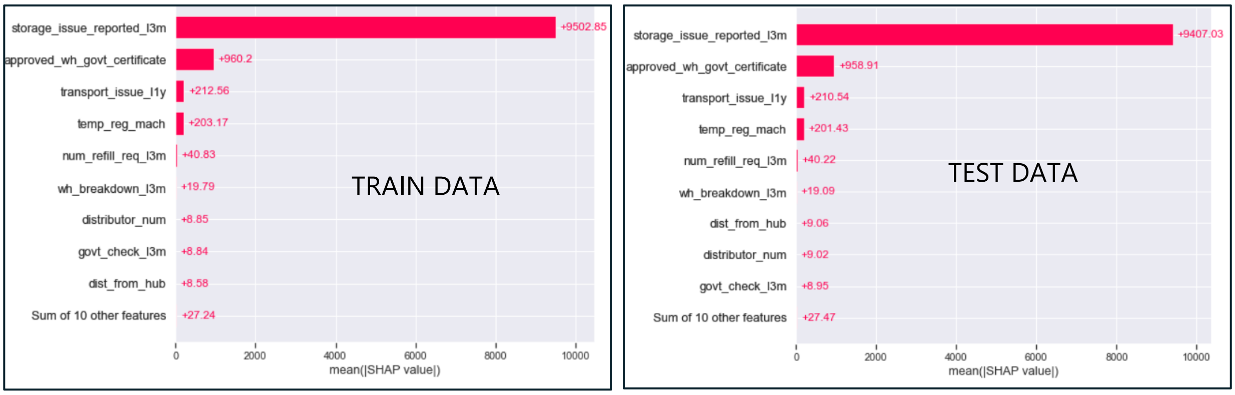


FIGURE 4-2: RANDOMFORESTREGRESSOR FEATURE IMPORTANCE PLOT FOR TRAIN & TEST

4.b.4 - XGBRegressor

XGBRegressor fitted on Train dataset and the results are shown below.

R-Squared (1- SSE/SST) Train data: 0.995

R-Squared (1- SSE/SST) Test data: 0.994

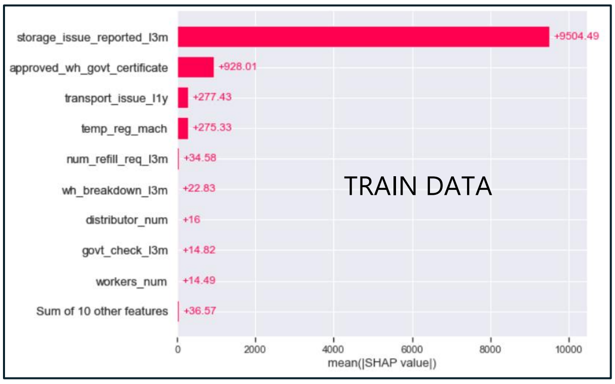
RMSE Train data: 852.986

RMSE Test data: 869.539

Mean Abs % Error Train data: 4.063 %

Mean Abs % Error Test data: 4.211 %

Feature Importance’s



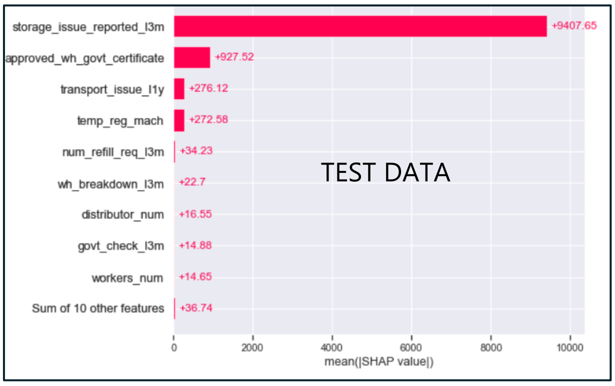


FIGURE 4-3: XGBREGRESSOR FEATURE IMPORTANCE PLOT FOR TRAIN & TEST

4.b.5 - MLPRegressor

MLPRegressor fitted on Train dataset and the results are shown below.

R-Squared (1- SSE/SST) Train data: 0.993

R-Squared (1- SSE/SST) Test data: 0.993

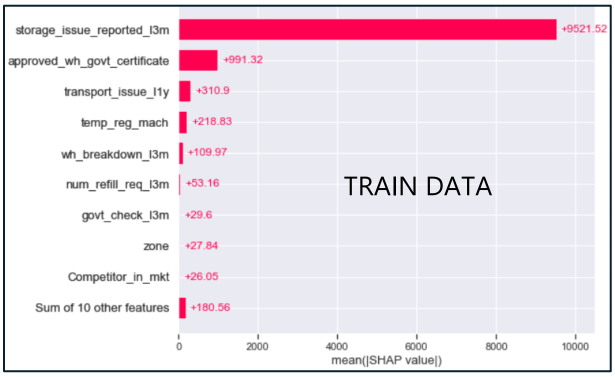
RMSE Train data: 943.381

RMSE Test data: 966.987

Mean Abs % Error Train data: 4.709 %

Mean Abs % Error Test data: 4.865 %

Feature Importance’s



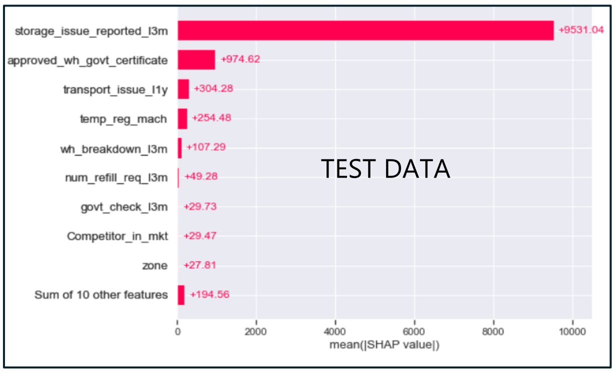


FIGURE 4-4: XGBREGRESSOR FEATURE IMPORTANCE PLOT FOR TRAIN & TEST

4.c - Model Tuning

4.c.1 - Ensemble modelling

Ensemble modelling is used of two types.

- Bagging
- Boosting

for Bagging model Random Forest regressor is used and

for Boosting case XGB regressor is used.

Refer section ***RandomForestRegressor***, ***XGBRegressor*** for more details.

4.c.2 - model tuning (to improve performance)

Hyper parameter & cross validation is done simultaneously using GridSearchCV, which can parameter tuned as well as cross validated for high model performance.

Below are the models where parameter tuning is done.

- **RandomForestRegressor**
  Parameters tuned were max_depth, n_estimators, max_features, min_samples_split and final parameter values shown below

```
'n_estimators'      : [100],
"max_features"      : ["auto"],
'max_depth'         : [9],
'min_samples_split' : [3],
'min_samples_leaf'  : [3],
'criterion'         : ['squared_error']
```

- **XGBRegressor**
  Parameters tuned were max_depth, n_estimators, reg_alpha, reg_lambda, learning_rate and final parameter values shown below:

```
'learning_rate': [0.1],
'n_estimators': [100],
'gamma': [0],
'objective' : ['reg:squarederror'],
'max_depth': [5], # -1.0
'reg_alpha': [0.8]
'reg_lambda': [1.2]
'colsample_bytree' : [1],
'subsample': [1],
'min_child_weight': [0.001],
```

- **MLPRegressor**
  Parameters tuned were max_ hidden_layer_sizes, activation, max_iter and final parameter values shown below:
  'hidden_layer_sizes' :[(100,100,100)],
  'activation' :['relu'],
  'solver' :['lbfgs'],
  'max_iter' :[200],

Parameter tuning definitely improved R-Square metrics significantly.

storage_issue_reported_l3m attribute is the most important attribute with more than 95% of importance in all the attributes, which is found in all the models.

Sequential feature selection is performed to find most important features/columns/attributes for score improvement, it is found that inclusion of all features did not increase or decrease the score. Hence in the final model all features were included.

5 - MODEL VALIDATION

5.a - Approach

- Model validation is done by using metrics like R-Squared, Root mean square error(RMSE), Mean absolute percentage error(MAPE).
- The entire dataset is split into train data→ 70% of random data & test data→ 30% remaining data.
- Model is built on Train dataset & is validated on test dataset using metrics like R-squared, RMSE & MAPE.
- To ensure model performance hyper parameter tuning & cross validation has been done using GridSearchCV.

5.b - Validation of test dataset

Linear regression Metrics

Linear Regression	Train	Test	%_change
R-Squared (1- SSE/SST)	0.977	0.978	0.10%
RMSE	1748.132	1692.693	3.17%
Mean Abs % Error	8.80%	8.69%	1.27%

Random Forest Regressor Metrics

Random Forest Regressor	Train	Test	%_change
R-Squared (1- SSE/SST)	0.995	0.994	0.10%
RMSE	852.301	893.355	4.82%
Mean Abs % Error	4.00%	4.22%	5.63%

XGB Regressor Metrics

XGB Regressor	Train	Test	%_change
R-Squared (1- SSE/SST)	0.995	0.994	0.10%
RMSE	852.986	869.539	1.94%
Mean Abs % Error	4.06%	4.21%	3.64%

MLP Regressor Metrics

MLP Regressor	Train	Test	%_change
R-Squared (1- SSE/SST)	0.993	0.993	0.00%
RMSE	943.381	966.987	2.50%
Mean Abs % Error	4.71%	4.87%	3.31%

- All models performed above par with R-Squared more than 97%.
- XGB has the highest metrics score with close fitting on train over test dataset.
- MLP regressor almost perfectly fit but consumes most amount of time while building.

5.c - Model comparison

ML Model	Metrics	TRAIN	TEST	fit_%
Linear Regression	R-Squared	0.977	0.978	0.10%
Random Forest Regressor	R-Squared	0.995	0.994	0.10%
XGBoost Regressor	**R-Squared**	**0.995**	**0.994**	**0.10%**
MLP Regressor	**R-Squared**	**0.993**	**0.993**	**0.00%**
ML Model	**Metrics**	**TRAIN**	**TEST**	**fit_%**
Linear Regression	RMSE	1748.132	1692.693	3.17%
Random Forest Regressor	RMSE	852.301	893.355	4.82%
XGBoost Regressor	**RMSE**	**852.986**	**869.539**	**1.94%**
MLP Regressor	**RMSE**	**941.458**	**970.665**	**3.10%**
ML Model	**Metrics**	**TRAIN**	**TEST**	**fit_%**
Linear Regression	MAPE	8.80%	8.69%	1.27%
Random Forest Regressor	MAPE	4.00%	4.22%	5.63%
XGBoost Regressor	**MAPE**	**4.06%**	**4.21%**	**3.64%**
MLP Regressor	**MAPE**	**4.68%**	**4.87%**	**3.93%**

5.d - Observations

- XGB Regressor outperforms in all the metrics when compared to other models.
- It also has very low under/Over fitting.
- MLP regressor has performed well in comparison with perfect fitting of model.
- MLP Regressor consumes more time than XGB Regressor.

5.e - Interpretation of the model(s)

- Random forest & XGB regressor outperformed in model fitting with more than 99.5% and MLP regressor is very close to top two models and fitted perfectly.
- XGB regressor & Random Forest outperformed with least RMSE values. But Random Forest has been overfitted.
- XGB regressor & Random Forest outperformed with least MAPE values. But Random Forest has been overfitted. MLP is close with good fitting of model.
- XGB regressor outperformed in every aspect of metric.
- **storage_issue_reported_l3m** attribute is the most important attribute with more than 95% of importance in all the attributes.

6 - FINAL INTERPRETATION/RECOMENDATION

6.a - Final Model

XGBRegressor is chosen as the most optimized final model for the following reasons.

- Highest R-square score of all models with more than 99.5%.
- Least RMSE & MAPE metrics of all models.
- Least in under/Over fitting when compared to other models in all the metrics.
- Although MLP Regressor is very close to XGB Regressor, it is observed that MLP Regressor consumed more significant time than XGB.

6.b - Business Implications

- Storage issue reported is the sole culprit in the entire attributes, which is significantly affecting the product weight that need to be sent to warehouse.
- Storage issues like rats, fugus due to moisture will obviously affect the product, which is instant noodles, and is leading to damage of noodle packages.
- So, the conclusion is we are having an **Inventory Loss**.

6.c - Insights from analysis

- XGB Regressor is the final model chosen for highest in metrics calculation, least in Under/Over fitting & least time consumed when compared to other models.
- ***Storage Issue reported*** in the last three months like rats' issue, fungus, moisture related etc is the **most important feature** which is **contributing** more than **98%** for the **model**.
- From below figure it is clear that, as number of storage issues increases average weight of product sent to warehouse also increases.

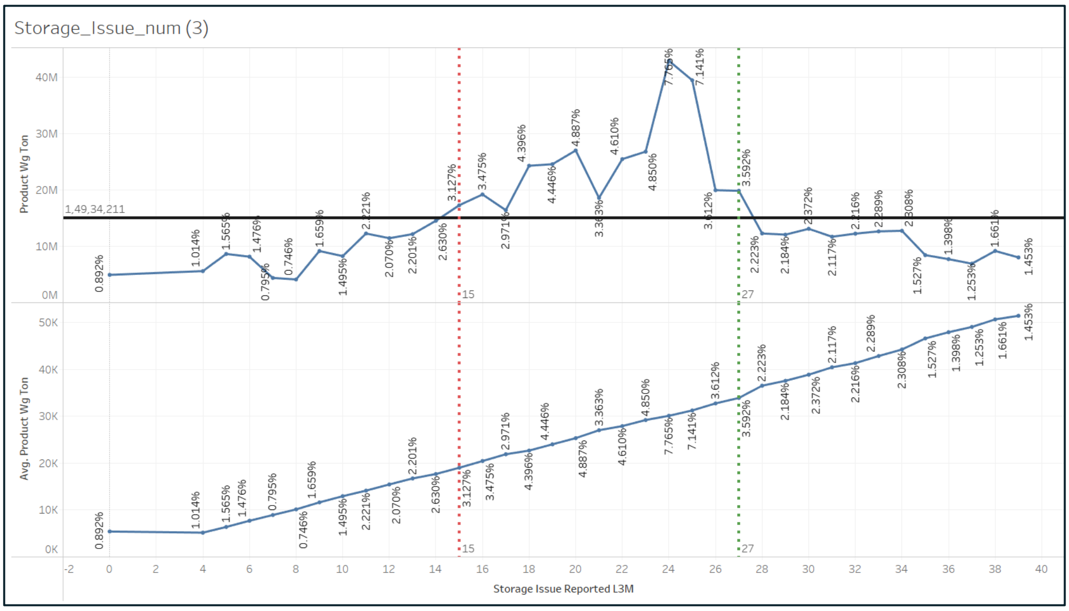


FIGURE 6-1: AVERAGE PRODUCT WEIGHT VS STORAGE ISSUE

- Storage issues between 15 to 27 need to be checked if it is overloading as these are punching above average.

6.d - Business Recommendations

- Top Priority to reduce number of Storage Issues, which is affecting the product weight significantly.
- Machine sensors for detecting temperature, moisture & potential pests.
- Designing a proper Architecture of Warehouse based on current Storage Issues.
- Assurance of Safety Standards, Health & Hygiene of Warehouse.
- More specific data of storage is to be collected like type, magnitude & frequency of issues etc.
- Run XGBoost model to predict optimize weight for given storage issues.
- Target Rural areas which are 90% of revenue.
- Target North/West Zone5,6 which is 50% of revenue.