Name: Kuntumalla Jayashree
Date: 21/05/2023

CAPSTONE PROJECT

FMCG COMPANY DATA

TABLE OF CONTENTS

List of Tables

Table1: Data Dictionary.....4

List of figures

Fig1: Data Sample.....4
Fig2: Data Info.....4
Fig3: Descriptive statistics.....5
Fig4: Mode.....5
Fig5: Countplot-Categorical.....6
Fig6: Distribution plot-Numerical.....7
Fig7: Bivariate-Analysis-Categorical.....8
Fig8: Bivariate-Analysis-Numerical.....9
Fig9: Correlation.....10
Fig10: Pair plot.....11
Fig11: Null values.....11
Fig12: Outliers.....12
Fig13: Sample data-Transformed.....12
Fig14: Corr-Features.....13
Fig15: SQRT transformation plot.....13
Fig16: Yeojohnson transformation plot.....14
Fig17: Cluster info.....15
Fig18: Summary-OLS.....16
Fig19: VIF1.....17
Fig20: Summary_cluster0_after_dropping_dummy.....17
Fig21: VIF_cluster0_after_dropping_dummy.....18
Fig22: Final_summary_cluster0.....18
Fig23: Fitted Vs Residual plot.....18
Fig24: Normality of Residuals.....19
Fig25: Q-Q plot.....19
Fig26: Final Model Summary+Cluster0.....19
Fig27: Summary+Cluster1.....20
Fig28: VIF+Cluster1.....20
Fig29: After dropping dummy+Summary_Cluster1.....21
Fig30: Summary-After dropping non-Significant-Cluster1.....21
Fig31: VIF-After dropping non-Significant-Cluster1.....21
Fig32: Fitted Vs Residual plot-cluster1.....22
Fig33: Normality of Residuals-cluster1.....22
Fig34: Q-Q plot-cluster1.....22
Fig35: Final Model-Cluster1.....23
Fig36: Models Comparison-before tuning.....23
Fig37: Feature importance as per gradient boosting.....24
Fig38: Model comparison after tuning.....24
Fig39: Feature imp-plot-GBR-1.....25
Fig40: Feature imp-plot-GBR-0.....25

Questions

Introduction - What did you wish to achieve while doing the project ?	3
EDA - Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.	6
Data Cleaning and Pre-processing - Approach used for identifying and treating missing values and outlier treatment (and why) - Need for variable transformation (if any) - Variables removed or added and why (if any)	11
Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.	15
Model validation - How was the model validated ? Just accuracy, or anything else too ?	15
Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client.	25

Introduction of the business problem:

Problem Statement:

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

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

Also try to analysis the demand pattern in different pockets of the country so management can drive the advertisement campaign particular in those pockets. This is the first phase of the agreement; hence, company has shared very limited information. Once you are able to showcase a tangible impact with this much of information then company will open the 360 degree data lake for your consulting company to build a more robust model.

Need of this project is to identify what factors are causing the mismatch in the demand and the supply for a warehouse which is leading to inventory or business loss in-order to optimize the weight of the product shipped and to understand the demand pattern in different countries.

Data Dictionary:

Variable	Description
product_wg_ton	Product has been shipped in last 3 months. Weight is in tons
Ware_house_ID	Product warehouse ID
WH_Manager_ID	Employee ID of warehouse manager
Location_type	Location of warehouse like in city or village
WH_capacity_size	Storage capacity size of the warehouse
zone	Zone of the warehouse
WH_regional_zone	Regional zone of the warehouse under each zone
num_refill_req_l3m	Number of times refilling has been done in last 3 months
transport_issue_l1y	Any transport issue like accident or goods stolen reported in last one year
Competitor_in_mkt	Number of instant noodles competitor in the market
retail_shop_num	Number of retails shop who sell the product under the warehouse area
wh_owner_type	Company is owning the warehouse or they have get the warehouse on rent
distributor_num	Number of distributer works in between warehouse and retail shops
flood_impacted	Warehouse is in the Flood impacted area indicator
flood_proof	Warehouse is flood proof indicators. Like storage is at some height not directly on the ground
electric_supply	Warehouse have electric back up like generator, so they can run the warehouse in load shedding
dist_from_hub	Distance between warehouse to the production hub in Kms
workers_num	Number of workers working in the warehouse
wh_est_year	Warehouse established year
storage_issue_reported_l3m	Warehouse reported storage issue to corporate office in last 3 months. Like rat, fungus because of moisture etc.
temp_reg_mach	Warehouse have temperature regulating machine indicator

approved_wh_govt_certificate	What kind of standard certificate has been issued to the warehouse from government regulatory body
wh_breakdown_l3m	Number of time warehouse face a breakdown in last 3 months. Like strike from worker, flood, or electrical failure
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

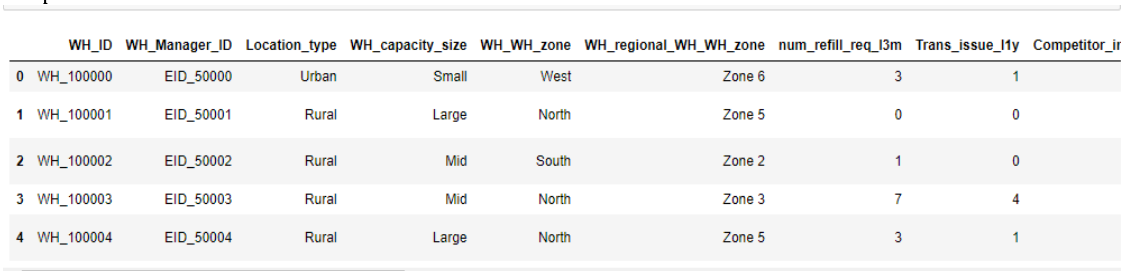
Table1: Data Dictionary

Data Report

- ➢ Load the required libraries, set the working directory and load the data file.
- ➢ The dataset has a total of 24 variables out of which 12 are categorical and rest is continuous variables.
- ➢ Changed column names as those are inappropriate.
- ➢ Changed the data type of flood_impacted, flood_proof, electric_supply and temp_reg_mach to object as these are defined as int but these are nominal variables.
- ➢ Shape (dimension) of the dataset is (25000, 24).
- ➢ There no duplicated values in the dataset.
- ➢ There are missing values in the dataset. Total 13779 missing values from workers num, wh est year and approved_wh_govt_certificate.

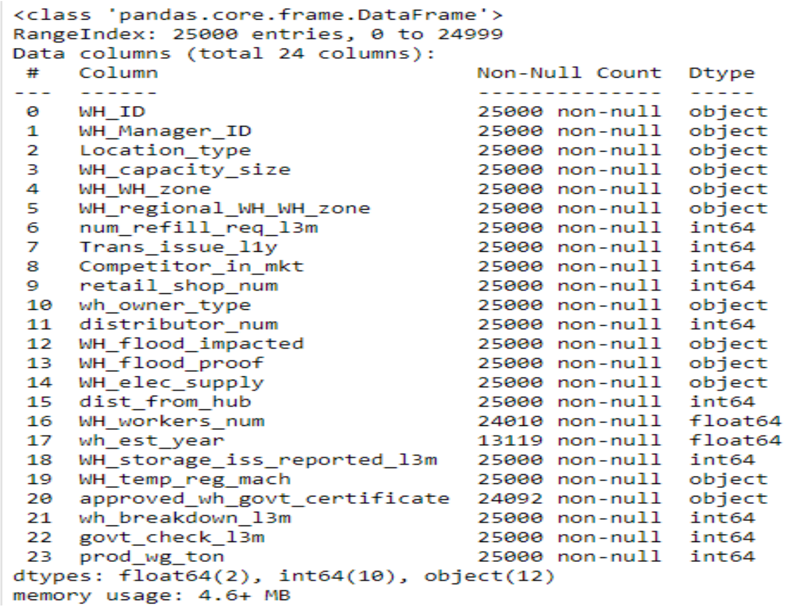
Let's start the data exploration with the head method of python.

Sample of the dataset:



	WH_ID	WH_Manager_ID	Location_type	WH_capacity_size	WH_WH_zone	WH_regional_WH_WH_zone	num_refill_req_l3m	Trans_issue_l1y	Competitor_ir
0	WH_100000	EID_50000	Urban	Small	West	Zone 6	3	1	
1	WH_100001	EID_50001	Rural	Large	North	Zone 5	0	0	
2	WH_100002	EID_50002	Rural	Mid	South	Zone 2	1	0	
3	WH_100003	EID_50003	Rural	Mid	North	Zone 3	7	4	
4	WH_100004	EID_50004	Rural	Large	North	Zone 5	3	1	

Fig1: Data Sample



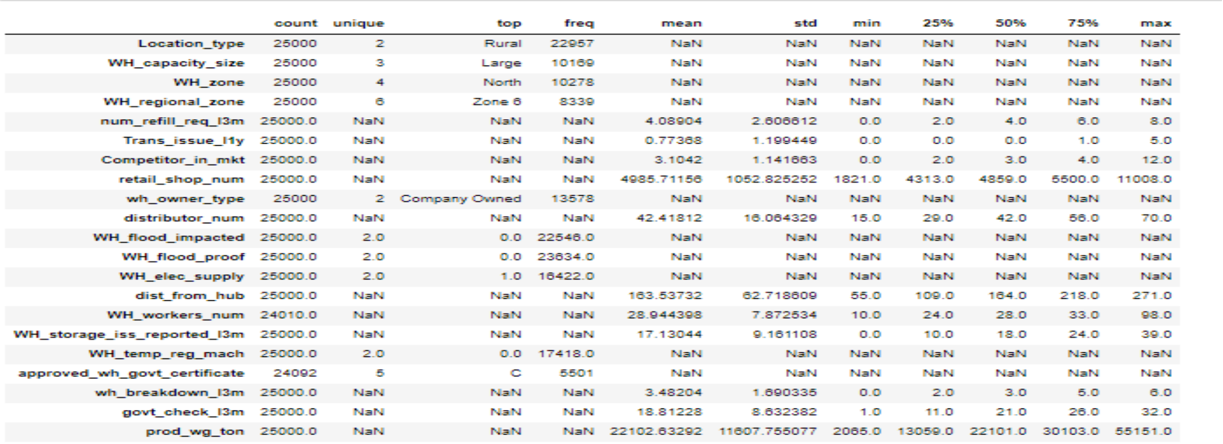
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   WH_ID                         25000 non-null  object 
 1   WH_Manager_ID                 25000 non-null  object 
 2   Location_type                 25000 non-null  object 
 3   WH_capacity_size              25000 non-null  object 
 4   WH_WH_zone                    25000 non-null  object 
 5   WH_regional_WH_WH_zone        25000 non-null  object 
 6   num_refill_req_l3m            25000 non-null  int64  
 7   Trans_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  WH_flood_impacted             25000 non-null  object 
 13  WH_flood_proof                25000 non-null  object 
 14  WH_elec_supply                25000 non-null  object 
 15  dist_from_hub                 25000 non-null  int64  
 16  WH_workers_num                24010 non-null  float64
 17  wh_est_year                   13119 non-null  float64
 18  WH_storage_iss_reported_l3m   25000 non-null  int64  
 19  WH_temp_reg_mach              25000 non-null  object 
 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  prod_wg_ton                   25000 non-null  int64  
dtypes: float64(2), int64(10), object(12)
memory usage: 4.6+ MB
```

Fig2: Data Info

Observations:

- ➢ Almost 47% of the data is missing in the year of establishment of warehouse variable.
- ➢ 86 % of the missing data is in the year of establishment of warehouse variable.
- ➢ As this variable with 47 % of missing info, won't give much addition to the analysis, we are dropping it for now.

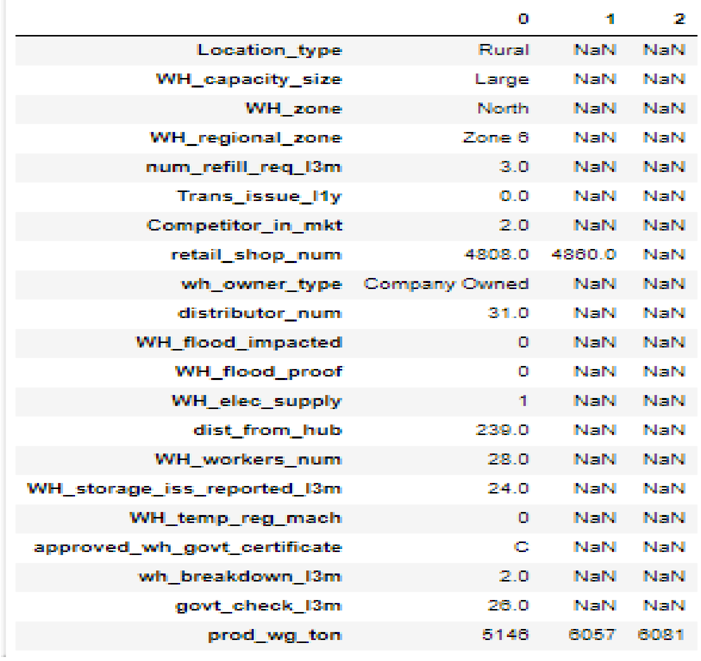
- We are dropping WH_ID and 'WH_Manager_ID as we don’t find any useful impact of these variables on the target variables.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Location_type	25000	2	Rural	22957	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_capacity_size	25000	3	Large	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_zone	25000	4	North	10278	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_regional_zone	25000	6	Zone 6	8339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num_refill_req_l3m	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
Trans_issue_l1y	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	NaN	NaN	NaN	4985.71156	1052.825252	1821.0	4313.0	4859.0	5500.0	11008.0
wh_owner_type	25000	2	Company Owned	13578	NaN	NaN	NaN	NaN	NaN	NaN	NaN
distributor_num	25000.0	NaN	NaN	NaN	42.41812	16.064329	15.0	29.0	42.0	56.0	70.0
WH_flood_impacted	25000.0	2.0	0.0	22546.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_flood_proof	25000.0	2.0	0.0	23634.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_elec_supply	25000.0	2.0	1.0	16422.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dist_from_hub	25000.0	NaN	NaN	NaN	163.53732	62.718609	55.0	109.0	164.0	218.0	271.0
WH_workers_num	24010.0	NaN	NaN	NaN	28.944398	7.872534	10.0	24.0	28.0	33.0	98.0
WH_storage_iss_reported_l3m	25000.0	NaN	NaN	NaN	17.13044	9.161108	0.0	10.0	18.0	24.0	39.0
WH_temp_reg_mach	25000.0	2.0	0.0	17418.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
approved_wh_govt_certificate	24092	5	C	5501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wh_breakdown_l3m	25000.0	NaN	NaN	NaN	3.48204	1.690335	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	NaN	NaN	NaN	18.81228	8.632382	1.0	11.0	21.0	26.0	32.0
prod_wg_ton	25000.0	NaN	NaN	NaN	22102.63292	11607.755077	2065.0	13059.0	22101.0	30103.0	55151.0

Fig3: Descriptive statistics

Mode:



	0	1	2
Location_type	Rural	NaN	NaN
WH_capacity_size	Large	NaN	NaN
WH_zone	North	NaN	NaN
WH_regional_zone	Zone 6	NaN	NaN
num_refill_req_l3m	3.0	NaN	NaN
Trans_issue_l1y	0.0	NaN	NaN
Competitor_in_mkt	2.0	NaN	NaN
retail_shop_num	4808.0	4860.0	NaN
wh_owner_type	Company Owned	NaN	NaN
distributor_num	31.0	NaN	NaN
WH_flood_impacted	0	NaN	NaN
WH_flood_proof	0	NaN	NaN
WH_elec_supply	1	NaN	NaN
dist_from_hub	239.0	NaN	NaN
WH_workers_num	28.0	NaN	NaN
WH_storage_iss_reported_l3m	24.0	NaN	NaN
WH_temp_reg_mach	0	NaN	NaN
approved_wh_govt_certificate	C	NaN	NaN
wh_breakdown_l3m	2.0	NaN	NaN
govt_check_l3m	26.0	NaN	NaN
prod_wg_ton	5146	6057	6081

Fig4: Mode

Observations:

- Location Type- rural area is where high number of the warehouses is established.
- Most of the warehouses are of large capacity.
- High numbers of warehouses are established in North Zone and Zone6.
- Highest number of refills done in the last 3 months is 8 while the average number of refills count is 4.08 with 4.0 as median.
- We observed that transportation issues raised in the last 1 year is on an average 0.77 times with 5 as maximum and 0 as median.
- Maximum number of competitors in the market is 12 with 3.1 as mean and 3 as median.
- There are around 11008 maximum retail shops with 4985.71 as average count.
- Maximum warehouses are owned by company, nearly 13578.
- Average distributor count is 42.4 with 70 as maximum.
- 22546 warehouses are not established in flood impacted area and so high number of them is no flood proof indicators. Along with these 22546 warehouses, there are extra 1088 warehouses that don’t possess flood proof indicator.
- 16422 warehouses have power back while rest doesn’t have.

- Average distance from hub is 163.53 kms with 271kms as maximum. While 50% of the warehouses are at a distance of 164kms from hub.
- On an average, there are 28.94 workers in the warehouses.
- Average Storage issues reported in the last 3 months is 17.1 while 50% of warehouses have 18 times.
- 17418 numbers of warehouses don't have temperature regulator indicator.
- C is the highest government approved certificate with a frequency of 5501.
- On an average there are 3.48 number of warehouse breakdowns in the last 3 months.
- Mean government check in the last 3 months is 18.8
- Average weight of the product is 22102.63.
- We see there are some missing values in the Workers number, approved_wh_govt_certificate.
- Almost all the variables are have mean and median as same with slight skewness.
- Highly skewed records are flood proof indicator ,flood affected areas, transport issue, workers count, temperature regulator indicator, electric supply, competitor in the market
- And retail shops count.

Exploratory data analysis

Univariate Analysis for Categorical Variables:

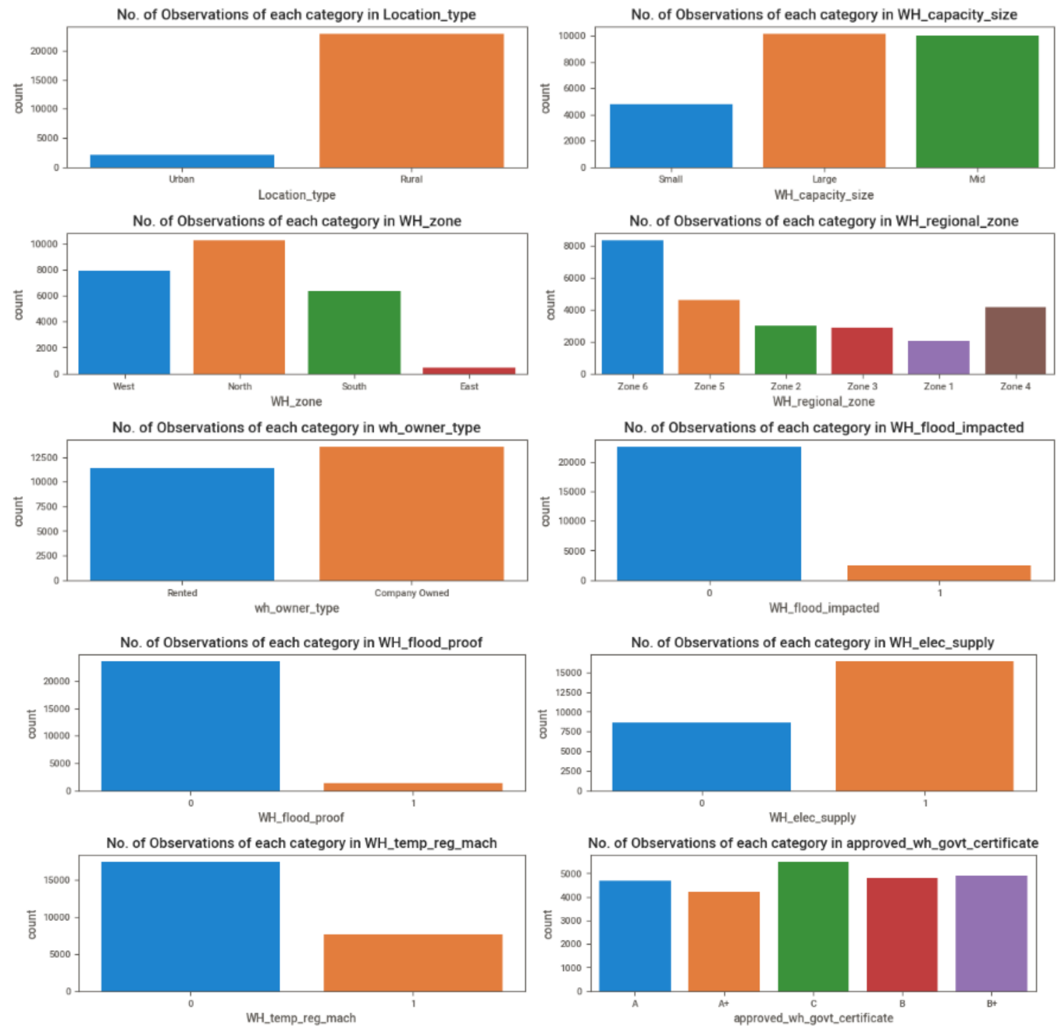


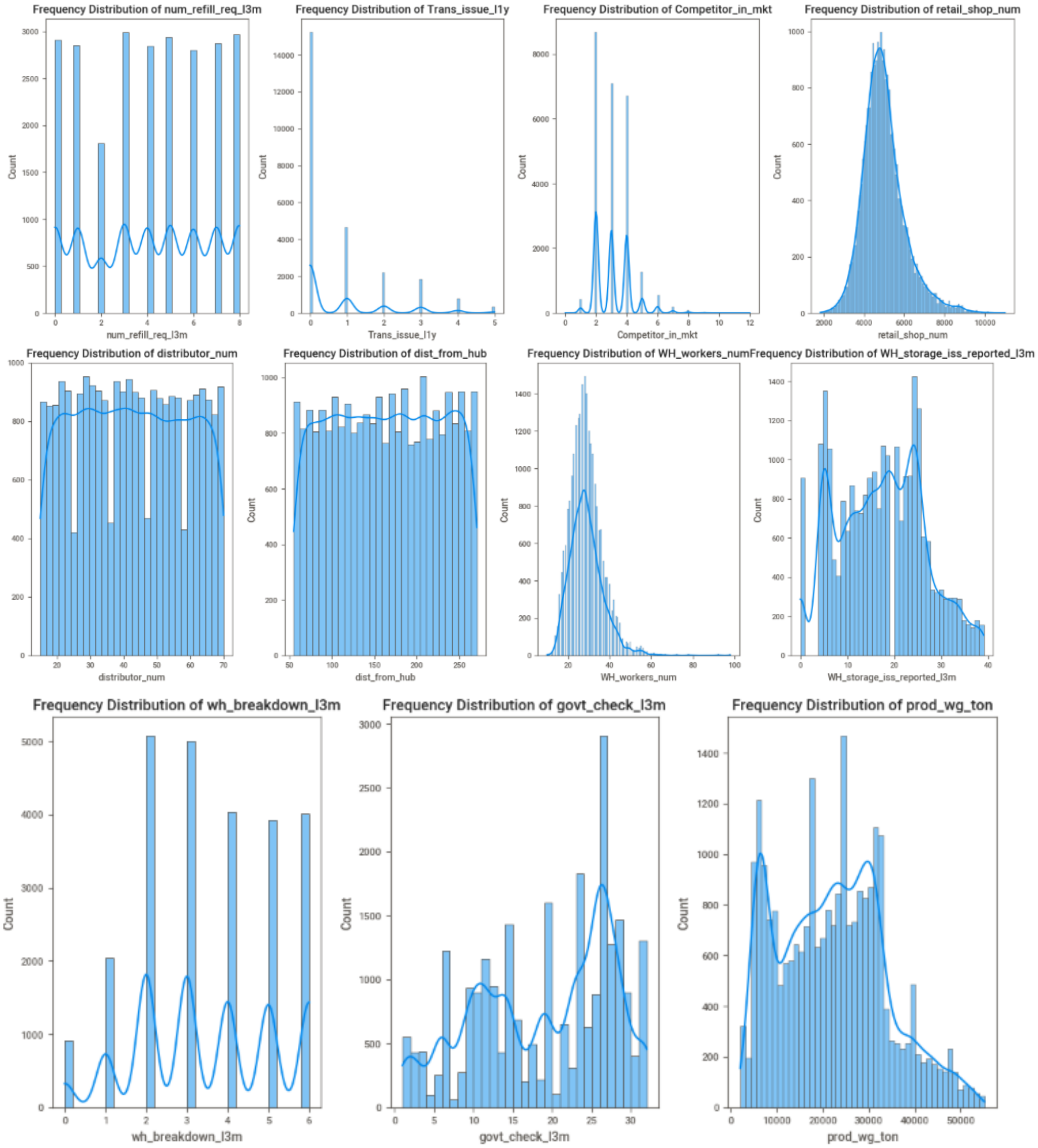
Fig5: Countplot-Categorical

Inferences:

- 91% of the warehouses are established in Rural areas.
- Nearly 80% of warehouses are Large and Mid size while only 20% is small size.
- North Zone warehouses constitute 41% while the East occupies only 1%.
- Zone6 Regional zone has 33% of warehouses establishment while Zone1 is the least with 8%.
- More than 50% of warehouses are owned by Company while 45% is rented.

- 90% of warehouses are not established in flood impacted areas and so most of them don’t have flood proof indicators.
- 65% of warehouses have power backups and nearly 70% of warehouses have temperature regulating machine indicator.
- Only around 35% of warehouses are of best ones while rest have some one or the other issues.

Univariate Analysis for Numerical Variables:



Frequency Distribution of num_refill_req_l3m

Frequency Distribution of Trans_issue_l1y

Frequency Distribution of Competitor_in_mkt

Frequency Distribution of retail_shop_num

Frequency Distribution of distributor_num

Frequency Distribution of dist_from_hub

Frequency Distribution of WH_workers_num

Frequency Distribution of WH_storage_iss_reported_l3m

Frequency Distribution of wh_breakdown_l3m

Frequency Distribution of govt_check_l3m

Frequency Distribution of prod_wg_ton

Fig6: Distribution plot-Numerical

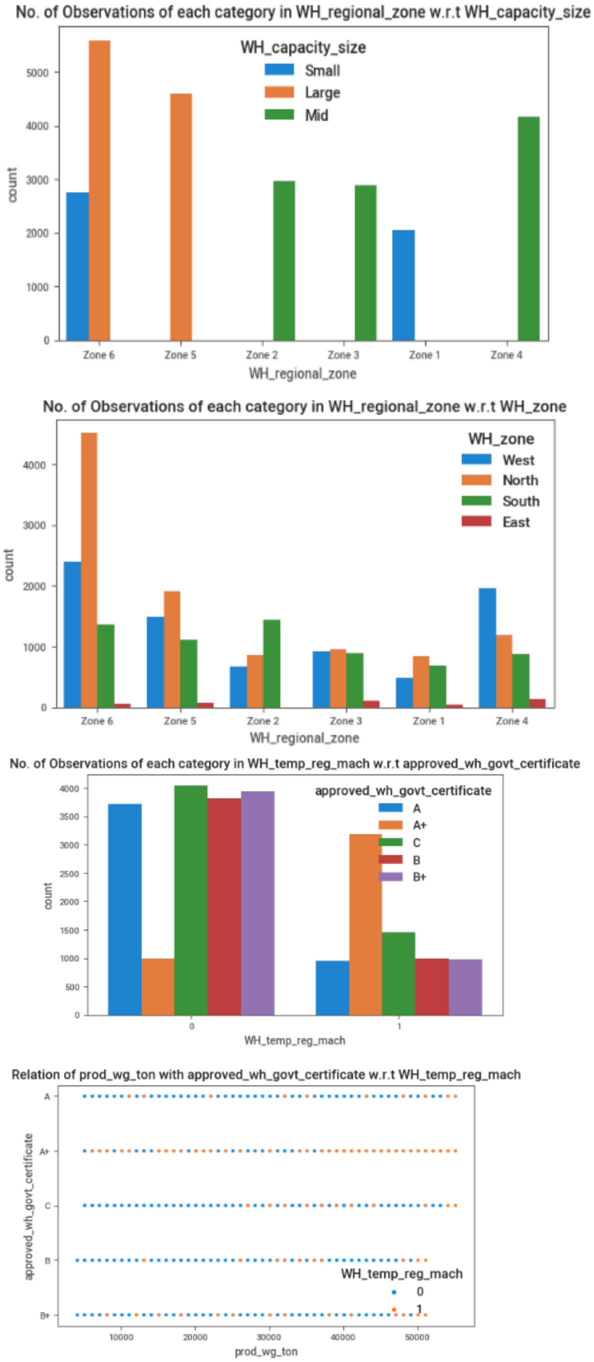
Observations:

- Mean and average of Target variable-product weight from descriptive stats is almost same which implies there is very less data skew.
- Most of the continuous variables or features are skewed as can observe mean and averages are not same.
- Overall retail shops count ranges from 2000 to 10000.
- Ware houses have workers count in a range from 10 to 60.
- Most of the ware houses breakdown at least 2-6 times
- Number of refills ranges from 0-8 times in the last 3 months.
- Government made 25-30 times warehouse checks in the last 3 months.
- Product weight is bimodal distribution.
- Maximum distance from the hub is around 300KMs.
- More Warehouses have less transportation issues

- ➢ Most of them have 2 competitors in the market.
- ➢ On an average 4-6 refills are done by most of the warehouses.

Bivariate Analysis :

- ➢ Taking Sweetviz visualisation report for exploration of relations between variables



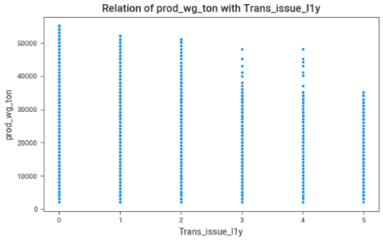
No. of Observations of each category in WH_regional_zone w.r.t WH_capacity_size

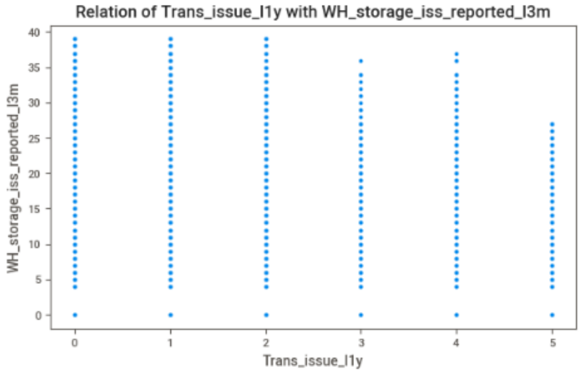
No. of Observations of each category in WH_regional_zone w.r.t WH_zone

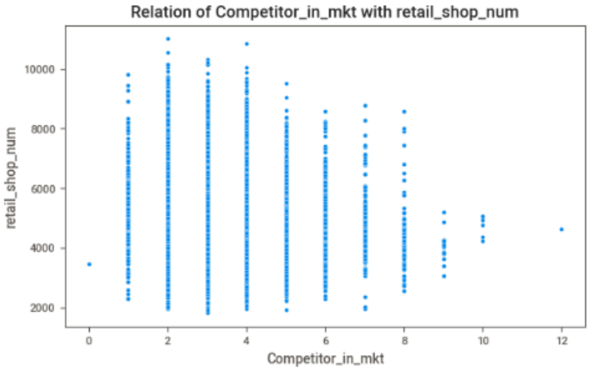
No. of Observations of each category in WH_temp_reg_mach w.r.t approved_wh_govt_certificate

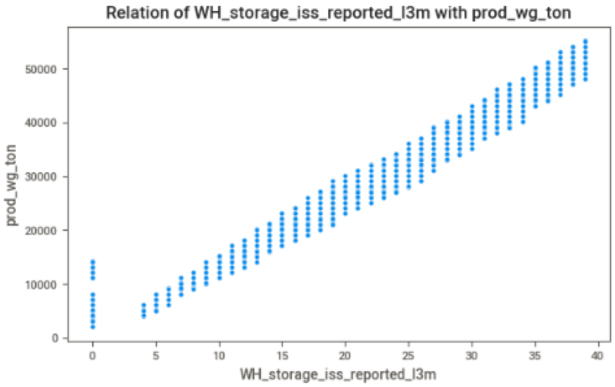
Relation of prod_wg_ton with approved_wh_govt_certificate w.r.t WH_temp_reg_mach

Fig7: Bivariate-Analysis-Categorical


Relation of prod_wg_ton with Trans_issue_l1y
prod_wg_ton
0
10000
20000
30000
40000
50000
0
1
2
3
4
5
Trans_issue_l1y


Relation of Trans_issue_l1y with WH_storage_iss_reported_l3m
WH_storage_iss_reported_l3m
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
Trans_issue_l1y


Relation of Competitor_in_mkt with retail_shop_num
retail_shop_num
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
Competitor_in_mkt


Relation of WH_storage_iss_reported_l3m with prod_wg_ton
prod_wg_ton
0
10000
20000
30000
40000
50000
0
5
10
15
20
25
30
35
40
WH_storage_iss_reported_l3m

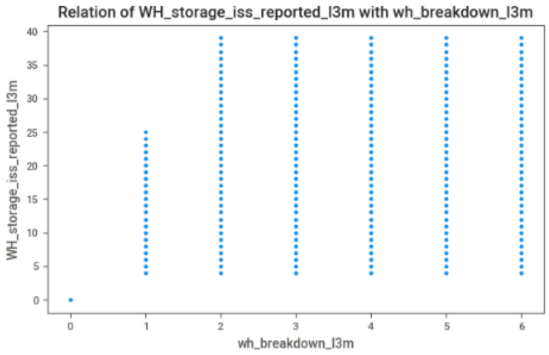
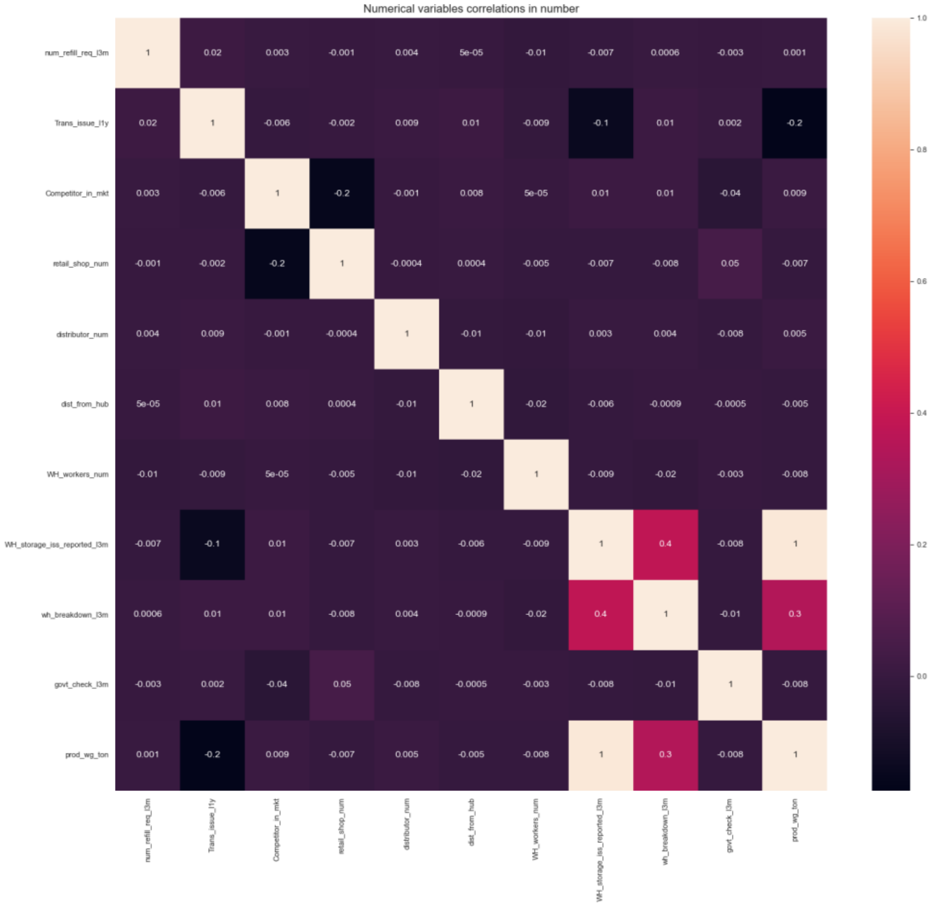


Fig8: Bivariate-Analysis-Numerical

Observations:

- Fewer breakdowns less storage issues.
- Increase in Storage issue reported increased the product weight.
- As competitors increases count of retail shop got decreased.
- When more Transport issues reported, storage issues reported got decreased.
- Product weight is reduced with increase in the transport issue in the last one year.
- Product weight is more with temp_reg and also has A+ certificate.

Correlation:



	num_refill_req_l3m	Trans_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	dist_from_hub	WH_workers_num	WH_storage_iss_reported_l3m	wh_breakdown_l3m	govt_check_l3m	prod_wg_ton
num_refill_req_l3m	1	0.02	0.003	-0.001	0.004	5e-05	-0.01	-0.007	0.0006	-0.003	0.001
Trans_issue_l1y	0.02	1	-0.006	-0.002	0.009	0.01	-0.009	-0.1	0.01	0.002	-0.2
Competitor_in_mkt	0.003	-0.006	1	-0.2	-0.001	0.008	5e-05	0.01	0.01	-0.04	0.009
retail_shop_num	-0.001	-0.002	-0.2	1	-0.0004	0.0004	-0.005	-0.007	-0.008	0.05	-0.007
distributor_num	0.004	0.009	-0.001	-0.0004	1	-0.01	-0.01	0.003	0.004	-0.008	0.005
dist_from_hub	5e-05	0.01	0.008	0.0004	-0.01	1	-0.02	-0.006	-0.0009	-0.0005	-0.005
WH_workers_num	-0.01	-0.009	5e-05	-0.005	-0.01	-0.02	1	-0.009	-0.02	-0.003	-0.008
WH_storage_iss_reported_l3m	-0.007	-0.1	0.01	-0.007	0.003	-0.006	-0.009	1	0.4	-0.008	1
wh_breakdown_l3m	0.0006	0.01	0.01	-0.008	0.004	-0.0009	-0.02	0.4	1	-0.01	0.3
govt_check_l3m	-0.003	0.002	-0.04	0.05	-0.008	-0.0005	-0.003	-0.008	-0.01	1	-0.008
prod_wg_ton	0.001	-0.2	0.009	-0.007	0.005	-0.005	-0.008	1	0.3	-0.008	1

Fig9: Correlation

Pair plots:

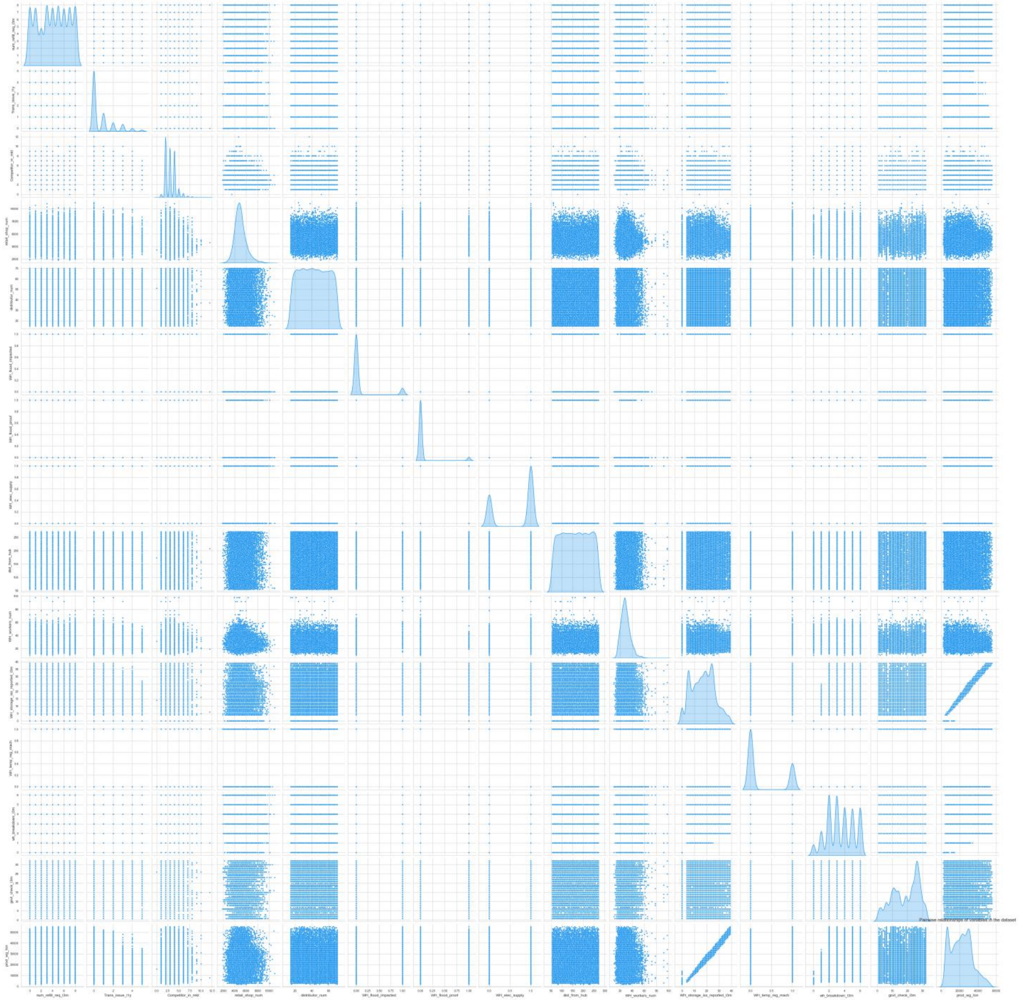


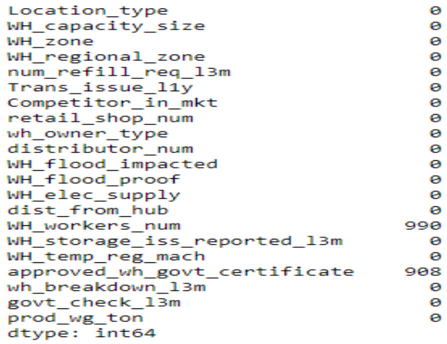
Fig10: Pair plot

Observations:

- Transportation issues are negatively correlated with warehouse issues and product weight.
- Retail shops number and competitors in the market are negatively related.
- Ware house break down and product weight are positively related.
- Ware house break down and storage issue are positively related.

Null Values Treatment:

- We observed there are missing values in WH_workers_num and approved_wh_govt_certificate.
- Using fillna() method , replaced missing values with mean in WH_workers_num and with mode in approved_wh_govt_certificate.



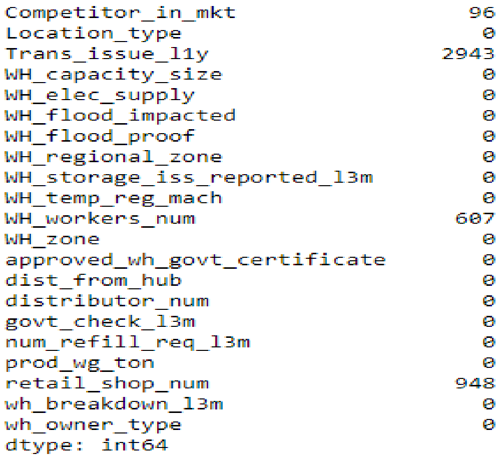
```
Location_type                     0
WH_capacity_size                  0
WH_zone                           0
WH_regional_zone                  0
num_refill_req_l3m                0
Trans_issue_l1y                   0
Competitor_in_mkt                 0
retail_shop_num                   0
wh_owner_type                     0
distributor_num                   0
WH_flood_impacted                 0
WH_flood_proof                    0
WH_elec_supply                    0
dist_from_hub                     0
WH_workers_num                  990
WH_storage_iss_reported_l3m       0
WH_temp_reg_mach                  0
approved_wh_govt_certificate    908
wh_breakdown_l3m                  0
govt_check_l3m                    0
prod_wg_ton                       0
dtype: int64
```

Fig11: Null values

Outliers:

We could see there are outliers in the Tran_iss_l1yr, Competitor_in_mkt, retail_shop_num, WH_workers_num
Using IQR-Inter Quartile range method, replaced all the data points that are greater than
Upper limit with upper limit and data points those are lower than the lower limit with lower limit.

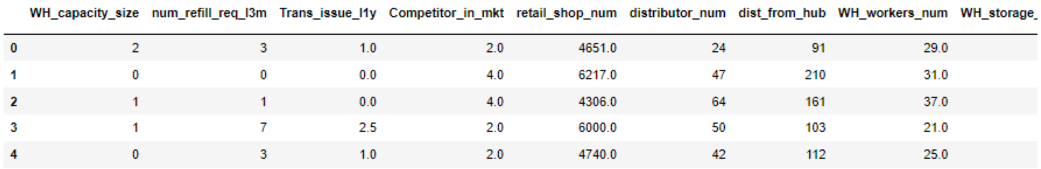
Q25 = 25 percentile
Q75= 75 percentile
IQR = Q75 – Q25



```
Competitor_in_mkt                 96
Location_type                      0
Trans_issue_l1y                 2943
WH_capacity_size                   0
WH_elec_supply                     0
WH_flood_impacted                  0
WH_flood_proof                     0
WH_regional_zone                   0
WH_storage_iss_reported_l3m        0
WH_temp_reg_mach                   0
WH_workers_num                   607
WH_zone                            0
approved_wh_govt_certificate       0
dist_from_hub                      0
distributor_num                    0
govt_check_l3m                     0
num_refill_req_l3m                 0
prod_wg_ton                        0
retail_shop_num                  948
wh_breakdown_l3m                   0
wh_owner_type                      0
dtype: int64
```

Fig12: Outliers

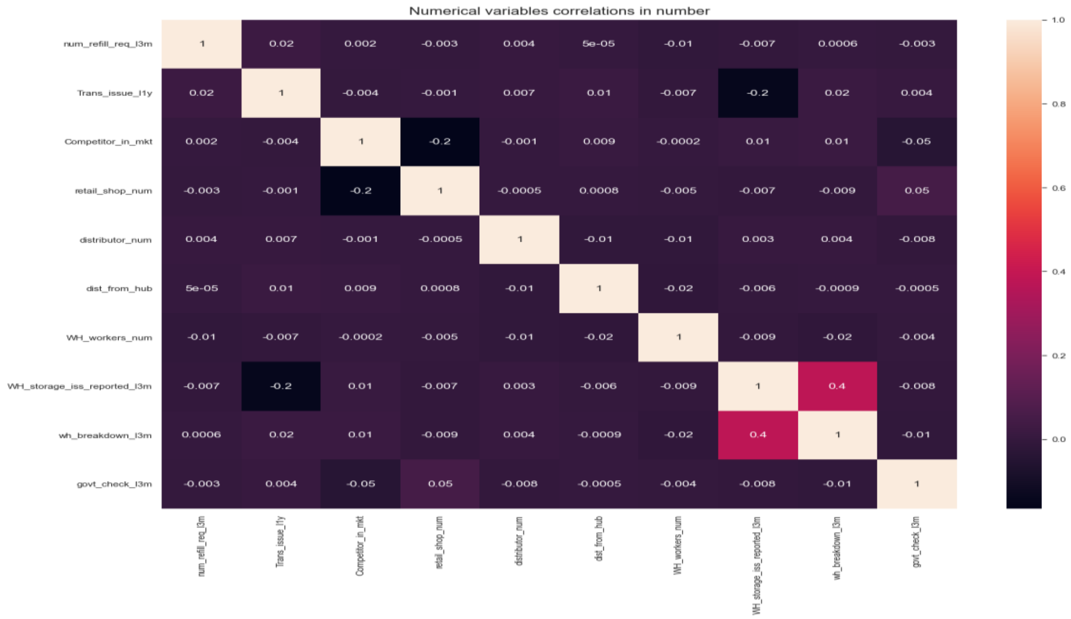
Transformed all nominal and categorical variables, using one hot encoding and label encoding for regression models. For all the variables that are below comes under the nominal encoding category so used label encoding for these variables transformation.
WH_regional_zone','WH_zone','wh_owner_type','WH_flood_impacted','WH_flood_proof',"WH_elec_supply, WH_temp_reg_mach, Location_type,WH_capacity_size and approved_wh_govt_certificate are ordinal variables so these are encoded using pd.categorical method and converted to 0's and 1's.



	WH_capacity_size	num_refill_req_l3m	Trans_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	dist_from_hub	WH_workers_num	WH_storage_
0	2	3	1.0	2.0	4651.0	24	91	29.0	
1	0	0	0.0	4.0	6217.0	47	210	31.0	
2	1	1	0.0	4.0	4306.0	64	161	37.0	
3	1	7	2.5	2.0	6000.0	50	103	21.0	
4	0	3	1.0	2.0	4740.0	42	112	25.0	

Fig13: Sample data-Transformed

Feature Selection:
Removal of Quasi-Constant features:
Quasi Constant Features are those that show the same value for a great majority of observations in the dataset. Here we have considered percentage of category in any variable to be greater than 99%. None of the categories in variables are greater than 99%.
Removal of duplicated features:
Often datasets contain duplicated features, that is, features that despite having different names are identical. In addition, we may often introduce duplicated features when performing one hot encoding of categorical variables, particularly if our datasets have many and /or highly cardinal categorical variables.
We have developed a function for identifying these features that are identical and found no feature is identical.
Removal of highly Correlated features:
Correlation matrix without target variable:



	num_refill_req_l3m	Trans_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	dist_from_hub	WH_workers_num	WH_storage_iss_reported_l3m	wh_breakdown_l3m	govt_check_l3m
num_refill_req_l3m	1	0.02	0.002	-0.003	0.004	5e-05	-0.01	-0.007	0.0006	-0.003
Trans_issue_l1y	0.02	1	-0.004	-0.001	0.007	0.01	-0.007	-0.2	0.02	0.004
Competitor_in_mkt	0.002	-0.004	1	-0.2	-0.001	0.009	-0.0002	0.01	0.01	-0.05
retail_shop_num	-0.003	-0.001	-0.2	1	-0.0005	0.0008	-0.005	-0.007	-0.009	0.05
distributor_num	0.004	0.007	-0.001	-0.0005	1	-0.01	-0.01	0.003	0.004	-0.008
dist_from_hub	5e-05	0.01	0.009	0.0008	-0.01	1	-0.02	-0.006	-0.0009	-0.0005
WH_workers_num	-0.01	-0.007	-0.0002	-0.005	-0.01	-0.02	1	-0.009	-0.02	-0.004
WH_storage_iss_reported_l3m	-0.007	-0.2	0.01	-0.007	0.003	-0.006	-0.009	1	0.4	-0.008
wh_breakdown_l3m	0.0006	0.02	0.01	-0.009	0.004	-0.0009	-0.02	0.4	1	-0.01
govt_check_l3m	-0.003	0.004	-0.05	0.05	-0.008	-0.0005	-0.004	-0.008	-0.01	1

Fig14: Corr-Features

Correlation coefficients whose magnitude is between 0.7 and 0.9 indicate variables which can be considered highly correlated.

There are no features that correlated with magnitude greater than 0.7(threshold).

Transformation of data:

We have observed there are 2 features that don't follow normality. They have skewness greater than 0.5 and less than -0.5.

Trans_issue_l1y and Competitor_in_mkt are the features that don't follow normality. So performing transformations on these features to convert them to normality. Since both the columns are having

0 in them, log transformations is not possible as this is applicable strictly to above 0 and reciprocal transformation also not possible as it needs non-zero values

Normality distribution after performing transformations:

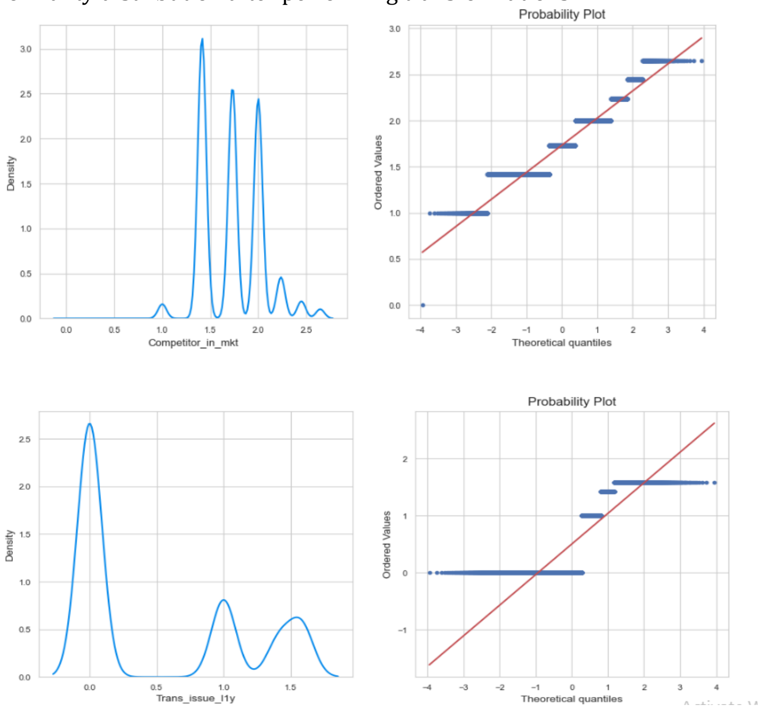


Fig15: SQRT transformation plot

Skew for Competitor in market variables after square root transformation is 0.317.Skew for Transport issue variables after square root transformation is 0.633

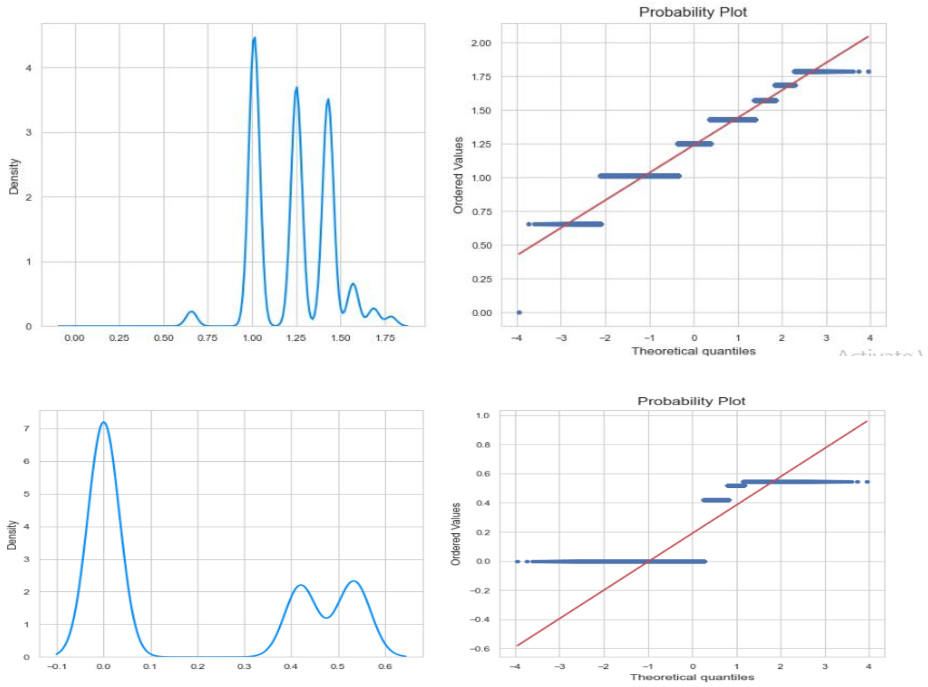
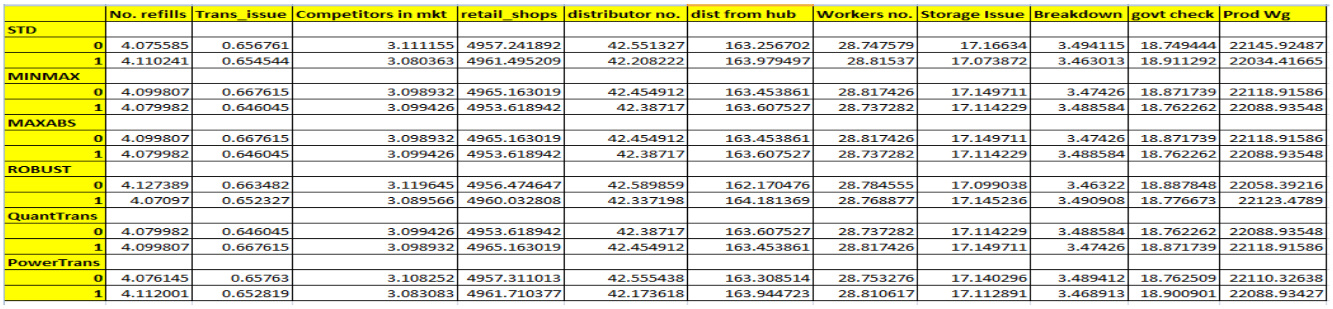


Fig16: Yeojohnson transformation plot

Skew for Competitor in market variables after yeojohnson transformation is 0.004. Skew for Transport issue variables after yeojohnson transformation is 0.50. From above both transformations, we see yeojohnson gave us the best results. Gave more normally distributed data. As a part of Feature Engineering we have done clustering on the Scaled PCA data and clustered the data into 0 and 1 category.Here we have used different scaling methods and developed models separately to get the better performing model.Scaling methods used here are as follows:

- StandardScaler: It is Standardization type of scaling that transforms data to have zero mean and 1 standard deviation. It assumes data is normally distributed within each feature and scales them such that the distribution centered to 0 and deviation to 1. If data is not normally distributed then this is not the best scaling method to use. X new = (X-mean)/std
- Min Max Scaling: It is Normalization type of scaling that transforms data between [0,1] or [-1,1]. This is used when standard deviation is small and distribution is not Gaussian. This scaling is sensitive to outliers.
- Max Abs Scaling: It is Normalization type of scaling that transforms data to its maximum absolute values. Effected by outliers.X scaled = x/max(x)
- Robust Scaler: It is Normalization type of scaling. This works well with data having outliers. This scaler removes the median and scales the data according to the quantile range. Centering and scaling are based on percentiles and are there not influenced by a few numbers of huge marginal outliers.
- Qunatile transformer Scaler: This normalization scaling uses quantile information. It transforms features to follow a uniform or normal distribution. This scaling reduces the impact of outliers.
- Power transformer scaler: This normalization scaling is used to make more gaussian like data. Useful for modeling issues related to the variability of a variable that is unequal the range where normality is desired. It stabilizes the variance and minimizes skewness.

We have applied PCA on all different scaled data and used these in clustering purpose.
We have used K-means clustering for clustering data. K-means clustering is an unsupervised learning algorithm that is used to solve the clustering problems. The objective is to minimize the sum of the distances between the data points and the cluster centriod, to identify the correct group each data point should belong to.In all different scaled data, 2 is the optimal "K" value as we have seen there is a significant drop in the Elbow plot from 1 to 2.
Below table is the average info on the clustering data:



	No. refills	Trans_issue	Competitors in mkt	retail_shops	distributor no.	dist from hub	Workers no.	Storage Issue	Breakdown	govt check	Prod Wg
STD											
0	4.075585	0.656761	3.111155	4957.241892	42.551327	163.256702	28.747579	17.16634	3.494115	18.749444	22145.92487
1	4.110241	0.654544	3.080363	4961.495209	42.208222	163.979497	28.81537	17.073872	3.463013	18.911292	22034.41665
MINMAX											
0	4.099807	0.667615	3.098932	4965.163019	42.454912	163.453861	28.817426	17.149711	3.47426	18.871739	22118.91586
1	4.079982	0.646045	3.099426	4953.618942	42.38717	163.607527	28.737282	17.114229	3.488584	18.762262	22088.93548
MAXABS											
0	4.099807	0.667615	3.098932	4965.163019	42.454912	163.453861	28.817426	17.149711	3.47426	18.871739	22118.91586
1	4.079982	0.646045	3.099426	4953.618942	42.38717	163.607527	28.737282	17.114229	3.488584	18.762262	22088.93548
ROBUST											
0	4.127389	0.663482	3.119645	4956.474647	42.589859	162.170476	28.784555	17.099038	3.46322	18.887848	22058.39216
1	4.07097	0.652327	3.089566	4960.032808	42.337198	164.181369	28.768877	17.145236	3.490908	18.776673	22123.4789
QuantTrans											
0	4.079982	0.646045	3.099426	4953.618942	42.38717	163.607527	28.737282	17.114229	3.488584	18.762262	22088.93548
1	4.099807	0.667615	3.098932	4965.163019	42.454912	163.453861	28.817426	17.149711	3.47426	18.871739	22118.91586
PowerTrans											
0	4.076145	0.65763	3.108252	4957.311013	42.555438	163.308514	28.753276	17.140296	3.489412	18.762509	22110.32638
1	4.112001	0.652819	3.083083	4961.710377	42.173618	163.944723	28.810617	17.112891	3.468913	18.900901	22088.93427

Fig17: Cluster info

Insights:
For both clusters almost most of the average details are same with slight difference.
For categorical variables, all values are same for both clusters as per mode.

Model building and interpretation, Model Tuning and business implication:

Since this is a regression problem, the following supervised learning algorithms are used.

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- MLPRegressor
- Lasso Regression
- Support Vector Regression.
- BaggingRegressor
- Adaboosting Regressor
- GradientBoostingRegressor

Data is split into Train and test with 70 percent as train data and 30 percent as test data.
TrainTestSplit method is used for splitting of data into 70:30.
Metrics Used for validation:
Since this is a regression analysis, R2,Mean Absolute error ,RMSE and Accuracy are used to test the performance of the model. Models are developed separately on different clusters.
MAE: Mean Absolute error - MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
MAE = sum of all absolute errors/total number of observations
RMSE: Mean squared error states that finding the squared difference between actual and predicted value. RMSE is square root of mean squared error.
R2: R2 score is a metric that tells the performance of your model. It calculates how must regression line is better than a mean line.
R2 = 1 – squared sum error of regression line/squared sum error of mean line
Adjusted R square changes when variable is added or dropped but only when that variable is adding some value where as r square changes when any variables is added or dropped as there exist small amount of correlation between target and predictor variables and that change is chance correlation.
Accuracy: It refers how close a measurement is to the true or accepted value
Constructing Linear regression model on different scaled data for different clusters in-order to check which scaled data gives the best on both different clusters.
Cluster0:
Applying Linear regression using scikit learn and constructed a model using the train dataset.
Below are the correlation coefficients of the variables :

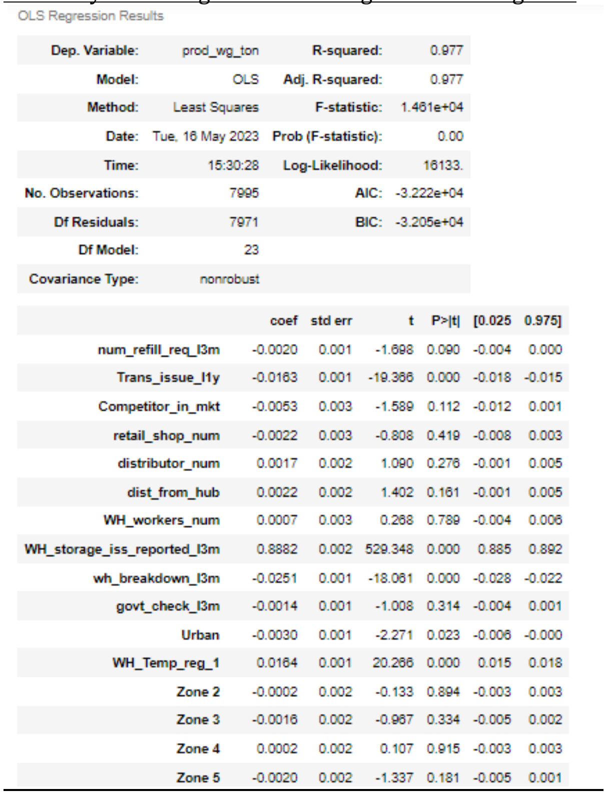
```
array([-2.05080652e-03, -1.63253659e-02, -5.39576538e-03, -2.46372502e-03,
        1.89308179e-03,  2.10124081e-03,  7.28469540e-04,  8.88221729e-01,
       -2.51664335e-02, -1.48059229e-03, -3.65937189e+10, -3.65937189e+10,
        1.11542194e+11,  1.11542194e+11, -2.42793718e+11, -2.42793718e+11,
       -2.42793718e+11, -2.42793718e+11, -2.42793718e+11, -2.42793718e+11,
       -2.81974137e+11, -2.81974137e+11, -2.81974137e+11, -2.81974137e+11,
        1.22070312e-04,  6.10351562e-05,  1.35534999e+11,  1.35534999e+11,
       -4.14165946e+10, -4.14165946e+10,  1.72349761e+10,  1.72349761e+10])
```

R square on training data: 0.9768274675848749
Adjusted R square on training data: 0.9767343350757963

R square on test data: 0.9771715140773358
Adjusted R square on test data: 0.9766089876092462
RMSE on training data: 0.03216898353306427
RMSE on testing data: 0.03182508757821573
Linear Regression Assumptions:
1. Independent and dependent variables are linearly related.
2. Residuals are independent
3. Equal variance of residuals
4. Residuals are normally distributed.
5. No or little multi-co linearity in independent variables.
Above assumptions violation results in mislead or biased forecasts, Confidence intervals and insights drawn from the model.

<u>Linear Regression using statsmodels(OLS):</u>
Constructed a Regression model using OLS method from statsmodels.
OLS(ordinary least squares): This method is used to minimize the sum of squared errors and generate the approximate estimates of coefficients.

<u>Summary of the Regression model generated using OLS:</u>



OLS Regression Results

Dep. Variable:	prod_wg_ton	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	1.461e+04
Date:	Tue, 16 May 2023	Prob (F-statistic):	0.00
Time:	15:30:28	Log-Likelihood:	16133.
No. Observations:	7995	AIC:	-3.222e+04
Df Residuals:	7971	BIC:	-3.205e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
num_refill_req_l3m	-0.0020	0.001	-1.698	0.090	-0.004	0.000
Trans_issue_l1y	-0.0163	0.001	-19.366	0.000	-0.018	-0.015
Competitor_in_mkt	-0.0053	0.003	-1.589	0.112	-0.012	0.001
retail_shop_num	-0.0022	0.003	-0.808	0.419	-0.008	0.003
distributor_num	0.0017	0.002	1.090	0.276	-0.001	0.005
dist_from_hub	0.0022	0.002	1.402	0.161	-0.001	0.005
WH_workers_num	0.0007	0.003	0.268	0.789	-0.004	0.006
WH_storage_iss_reported_l3m	0.8882	0.002	529.348	0.000	0.885	0.892
wh_breakdown_l3m	-0.0251	0.001	-18.061	0.000	-0.028	-0.022
govt_check_l3m	-0.0014	0.001	-1.008	0.314	-0.004	0.001
Urban	-0.0030	0.001	-2.271	0.023	-0.006	-0.000
WH_Temp_reg_1	0.0164	0.001	20.266	0.000	0.015	0.018
Zone 2	-0.0002	0.002	-0.133	0.894	-0.003	0.003
Zone 3	-0.0016	0.002	-0.967	0.334	-0.005	0.002
Zone 4	0.0002	0.002	0.107	0.915	-0.003	0.003
Zone 5	-0.0020	0.002	-1.337	0.181	-0.005	0.001

Fig18: Summary-OLS

Interpretation of R-squared: The R-squared value tells us that our model can explain 97.7% of the variance in the training set.
Interpretation of Coefficients:
* The coefficients tell us how one unit change in X can affect y. The sign of the coefficient indicates if the relationship is positive or negative.
* In this data set, an increase of 1 unit of warehouse storage issue reported in last 3 months occurs with a 0.8882 increase in the product weight.
* Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
* When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.
Interpretation of p-values (P > |t|):
* For each predictor variable there is a null hypothesis and alternate hypothesis.
   - Null hypothesis : Predictor variable is not significant
   - Alternate hypothesis : Predictor variable is significant

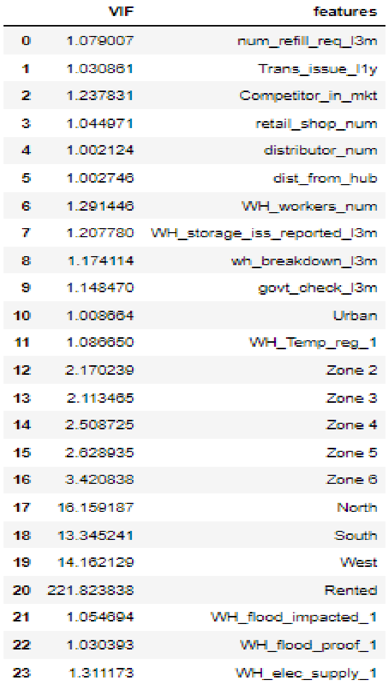
* (P > |t|) gives the p-value for each predictor variable to check the null hypothesis.
* If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.
* However, due to the presence of multicollinearity in our data, the p-values will also change.
* We need to ensure that there is no multicollinearity in order to interpret the p-values.

Multi- Co linearity:

Multi- Co linearity is checked by Variance Inflation factor. This tells what percentage of variance is inflated for each coefficient by the existence of correlation among the predictor variables in the model.

If VIF is 1, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of beta-k is not inflated at all.

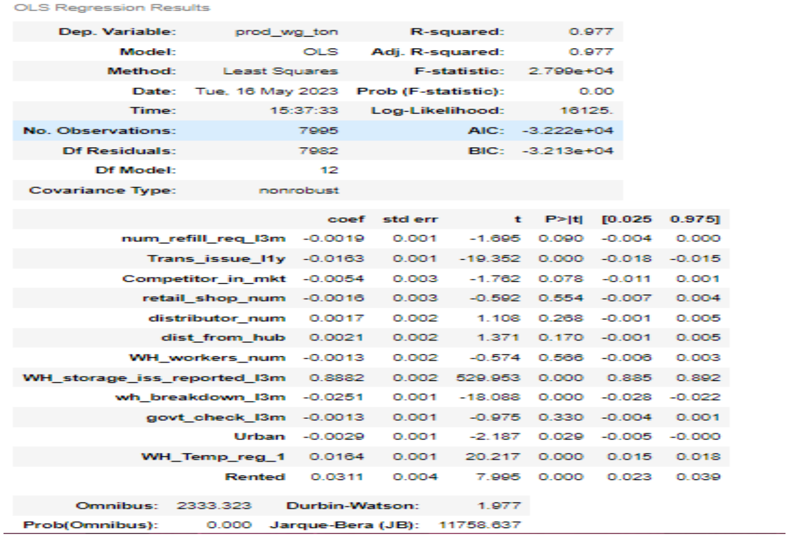
If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.



	VIF	features
0	1.079007	num_refill_req_l3m
1	1.030861	Trans_issue_l1y
2	1.237831	Competitor_in_mkt
3	1.044971	retail_shop_num
4	1.002124	distributor_num
5	1.002746	dist_from_hub
6	1.291446	WH_workers_num
7	1.207780	WH_storage_iss_reported_l3m
8	1.174114	wh_breakdown_l3m
9	1.148470	govt_check_l3m
10	1.008664	Urban
11	1.086650	WH_Temp_reg_1
12	2.170239	Zone 2
13	2.113465	Zone 3
14	2.508725	Zone 4
15	2.628935	Zone 5
16	3.420838	Zone 6
17	16.159187	North
18	13.345241	South
19	14.162129	West
20	221.823838	Rented
21	1.054694	WH_flood_impacted_1
22	1.030393	WH_flood_proof_1
23	1.311173	WH_elec_supply_1

Fig19: VIF1

VIF is not considered for categorical variables and so we are dropping dummy variables from model summary that have p>0.05 one by one.Final Summary after dropping Zone6,Zone2,Zone4,WH_flood_impacted_1,South,West,Zone 3,WH_flood_proof_1,WH_elec_supply_1,North,Zone 5 is as below:



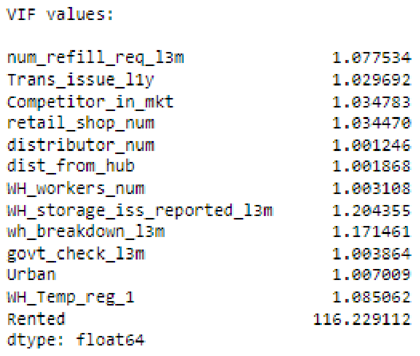
```
OLS Regression Results
      Dep. Variable:     prod_wg_ton          R-squared:      0.977
              Model:             OLS      Adj. R-squared:      0.977
             Method:   Least Squares          F-statistic:  2.799e+04
               Date: Tue, 16 May 2023  Prob (F-statistic):       0.00
               Time:        15:37:33      Log-Likelihood:     16125.
  No. Observations:            7995                  AIC: -3.222e+04
       Df Residuals:            7982                  BIC: -3.213e+04
           Df Model:              12
   Covariance Type:       nonrobust

                              coef  std err        t  P>|t|  [0.025  0.975]
        num_refill_req_l3m -0.0019    0.001   -1.695  0.090  -0.004   0.000
           Trans_issue_l1y -0.0163    0.001  -19.352  0.000  -0.018  -0.015
         Competitor_in_mkt -0.0054    0.003   -1.762  0.078  -0.011   0.001
           retail_shop_num -0.0016    0.003   -0.592  0.554  -0.007   0.004
           distributor_num  0.0017    0.002    1.108  0.268  -0.001   0.005
             dist_from_hub  0.0021    0.002    1.371  0.170  -0.001   0.005
            WH_workers_num -0.0013    0.002   -0.574  0.566  -0.006   0.003
WH_storage_iss_reported_l3m  0.8882    0.002  529.953  0.000   0.885   0.892
          wh_breakdown_l3m -0.0251    0.001  -18.088  0.000  -0.028  -0.022
            govt_check_l3m -0.0013    0.001   -0.975  0.330  -0.004   0.001
                     Urban -0.0029    0.001   -2.187  0.029  -0.005  -0.000
             WH_Temp_reg_1  0.0164    0.001   20.217  0.000   0.015   0.018
                    Rented  0.0311    0.004    7.995  0.000   0.023   0.039

        Omnibus:  2333.323      Durbin-Watson:      1.977
  Prob(Omnibus):     0.000  Jarque-Bera (JB):  11758.637
```

Fig20: Summary_cluster0_after_dropping_dummy

VIF after dropping:



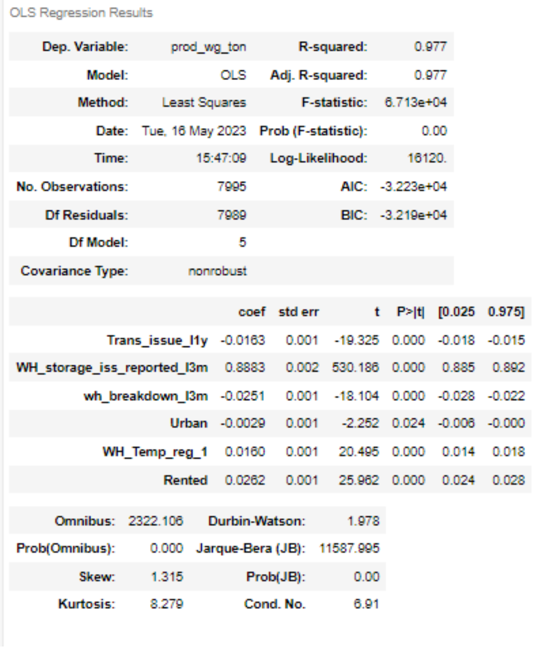
```
VIF values:

num_refill_req_l3m              1.077534
Trans_issue_l1y                 1.029692
Competitor_in_mkt               1.034783
retail_shop_num                 1.034470
distributor_num                 1.001246
dist_from_hub                   1.001868
WH_workers_num                  1.003108
WH_storage_iss_reported_l3m     1.204355
wh_breakdown_l3m                1.171461
govt_check_l3m                  1.003864
Urban                           1.007009
WH_Temp_reg_1                   1.085062
Rented                        116.229112
dtype: float64
```

Fig21: VIF_cluster0_after_dropping_dummy

In stats summary we see, there are variables whose p values are greater than 0.05 . These variables are not significant; we are dropping those variables one by one.

We have dropped all the variables that are not significant and developed final model as below



OLS Regression Results

Dep. Variable:	prod_wg_ton	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	6.713e+04
Date:	Tue, 16 May 2023	Prob (F-statistic):	0.00
Time:	15:47:09	Log-Likelihood:	16120.
No. Observations:	7995	AIC:	-3.223e+04
Df Residuals:	7989	BIC:	-3.219e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Trans_issue_l1y	-0.0163	0.001	-19.325	0.000	-0.018	-0.015
WH_storage_iss_reported_l3m	0.8883	0.002	530.186	0.000	0.885	0.892
wh_breakdown_l3m	-0.0251	0.001	-18.104	0.000	-0.028	-0.022
Urban	-0.0029	0.001	-2.252	0.024	-0.006	-0.000
WH_Temp_reg_1	0.0160	0.001	20.495	0.000	0.014	0.018
Rented	0.0262	0.001	25.962	0.000	0.024	0.028

Omnibus:	2322.106	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11587.995
Skew:	1.315	Prob(JB):	0.00
Kurtosis:	8.279	Cond. No.	6.91

Fig22: Final_summary_cluster0

Linearity Test:

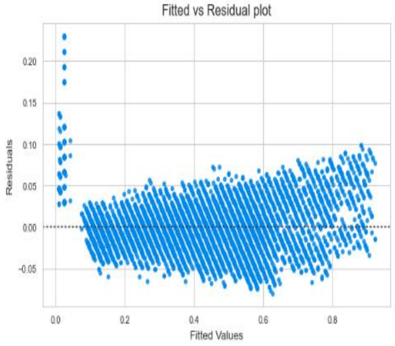


Fig23: Fitted Vs Residual plot

We can see that data is randomly distributed.

Normality Test:

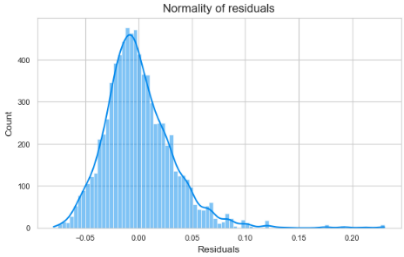


Fig24: Normality of Residuals

The residual terms are normally distributed.

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

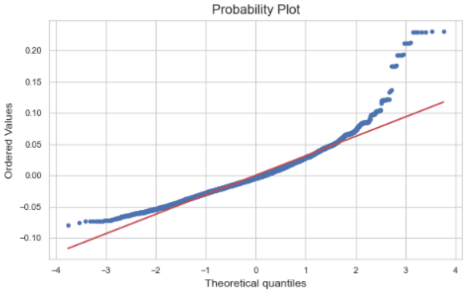


Fig25: Q-Q plot

Most of the points lie on the straight line.

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

* Null hypothesis - Data is normally distributed.
* Alternate hypothesis - Data is not normally distributed.

Shapiro test resulted in p value less than 0.05 , the residuals are not normally distributed as per shapiro test.

ShapiroResult(statistic=0.9353896379470825, pvalue=0.0)

TEST FOR HOMOSCEDASTICITY:

The null and alternate hypotheses of the goldfeldquandt test are as follows:

Null hypothesis : Residuals are homoscedastic

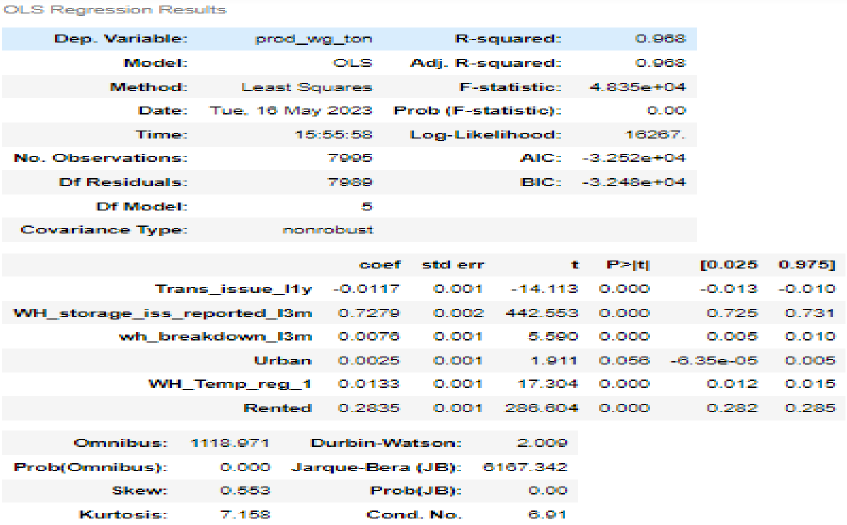
Alternate hypothesis : Residuals have hetroscedasticity

Result of test: [('F statistic', 1.062000646516132), ('p-value', 0.028718359788699747)]

p-value =0.02 < 0.05 , we can say that the residuals are not homoscedastic.

Since the homoscedastic is violated. We need to transform the target variable inorder to be homoscedastic.

Final Model:



```
OLS Regression Results

     Dep. Variable:     prod_wg_ton          R-squared:       0.968
             Model:             OLS      Adj. R-squared:       0.968
            Method:   Least Squares         F-statistic:   4.835e+04
              Date: Tue, 16 May 2023  Prob (F-statistic):       0.00
              Time:        15:55:58      Log-Likelihood:      16267.
 No. Observations:            7995                 AIC:  -3.252e+04
      Df Residuals:            7989                 BIC:  -3.248e+04
          Df Model:               5
 Covariance Type:        nonrobust

                              coef  std err        t  P>|t|    [0.025   0.975]
           Trans_issue_l1y  -0.0117    0.001  -14.113  0.000    -0.013   -0.010
WH_storage_iss_reported_l3m  0.7279    0.002  442.553  0.000     0.725    0.731
          wh_breakdown_l3m   0.0076    0.001    5.590  0.000     0.005    0.010
                     Urban   0.0025    0.001    1.911  0.056  -6.35e-05   0.005
             WH_Temp_reg_1   0.0133    0.001   17.304  0.000     0.012    0.015
                    Rented   0.2835    0.001  286.604  0.000     0.282    0.285

        Omnibus:  1118.971    Durbin-Watson:     2.009
 Prob(Omnibus):     0.000  Jarque-Bera (JB):  6167.342
           Skew:     0.553          Prob(JB):      0.00
       Kurtosis:     7.158          Cond. No.      6.91
```

Fig26: Final Model Summary+Cluster0

Observations:

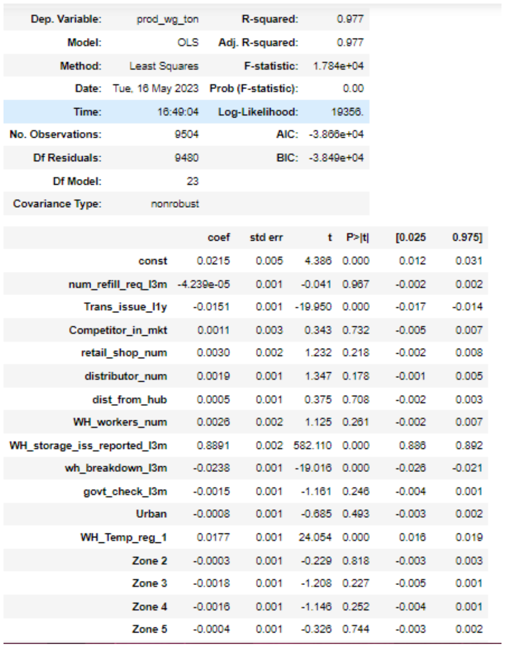
R-squared of the model is 0.968and adjusted R-squared is 0.968, which shows that the model is able to explain ~96% variance in the data. This is quite good. A unit increase in the warehouse storage issue reported in last 3 months will result in a 0.7279 unit increase in the product weight, all other variables remaining constant. A unit decrease in the transportation issue will result in a 0.0117 unit decrease in the product weight, all other variables remaining constant.

As per linear regression-cluster0, warehouse storage issue has the high importance while the ware house location in Urban has least importance among the important variables.

Model:

product weight = -0.01169080753664454 + 0.7278887860735657 * (WH_storage_iss_reported_l3m) + 0.007615153154223871 * (wh_breakdown_l3m) + 0.002454128078963348 * (Urban) + 0.013262900126872514 * (WH_Temp_reg_1) + 0.2834778818903032 * (Rented)

For Cluster1:



Dep. Variable:	prod_wg_ton	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	1.784e+04
Date:	Tue, 16 May 2023	Prob (F-statistic):	0.00
Time:	16:49:04	Log-Likelihood:	19356.
No. Observations:	9504	AIC:	-3.866e+04
Df Residuals:	9480	BIC:	-3.849e+04
Df Model:	23		
Covariance Type:	nonrobust		

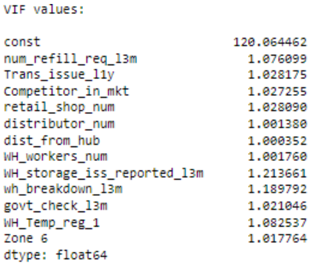
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0215	0.005	4.386	0.000	0.012	0.031
num_refill_req_l3m	-4.239e-05	0.001	-0.041	0.967	-0.002	0.002
Trans_issue_l1y	-0.0151	0.001	-19.950	0.000	-0.017	-0.014
Competitor_in_mkt	0.0011	0.003	0.343	0.732	-0.005	0.007
retail_shop_num	0.0030	0.002	1.232	0.218	-0.002	0.008
distributor_num	0.0019	0.001	1.347	0.178	-0.001	0.005
dist_from_hub	0.0005	0.001	0.375	0.708	-0.002	0.003
WH_workers_num	0.0026	0.002	1.125	0.261	-0.002	0.007
WH_storage_iss_reported_l3m	0.8891	0.002	582.110	0.000	0.886	0.892
wh_breakdown_l3m	-0.0238	0.001	-19.016	0.000	-0.026	-0.021
govt_check_l3m	-0.0015	0.001	-1.161	0.246	-0.004	0.001
Urban	-0.0008	0.001	-0.685	0.493	-0.003	0.002
WH_Temp_reg_1	0.0177	0.001	24.054	0.000	0.016	0.019
Zone 2	-0.0003	0.001	-0.229	0.818	-0.003	0.003
Zone 3	-0.0018	0.001	-1.208	0.227	-0.005	0.001
Zone 4	-0.0016	0.001	-1.146	0.252	-0.004	0.001
Zone 5	-0.0004	0.001	-0.326	0.744	-0.003	0.002

Fig27: Summary+Cluster1

VIF is not considered for categorical variables and so we are dropping dummy variables from model summary that have p>0.05 one by one.

After dropping those variables:

VIF:

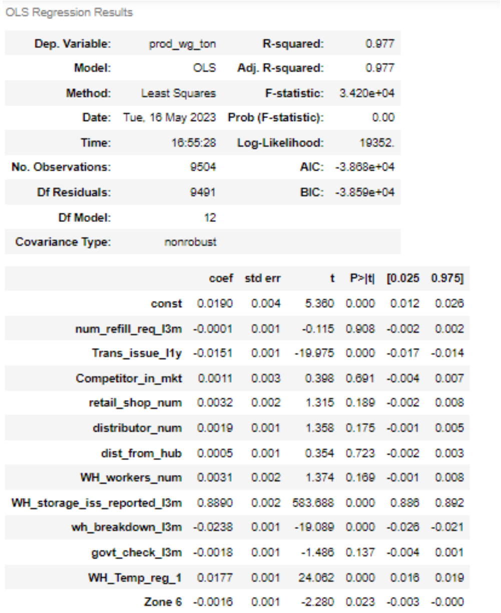


```
VIF values:

const                          120.064462
num_refill_req_l3m               1.076099
Trans_issue_l1y                  1.028175
Competitor_in_mkt                1.027255
retail_shop_num                  1.028090
distributor_num                  1.001380
dist_from_hub                    1.000352
WH_workers_num                   1.001760
WH_storage_iss_reported_l3m      1.213661
wh_breakdown_l3m                 1.189792
govt_check_l3m                   1.021046
WH_Temp_reg_1                    1.082537
Zone 6                           1.017764
dtype: float64
```

Fig28: VIF+Cluster1

Model Summary:



OLS Regression Results

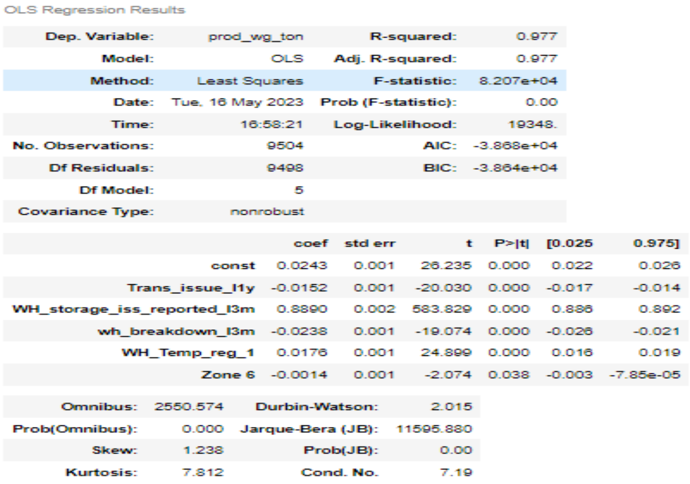
Dep. Variable:	prod_wg_ton	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	3.420e+04
Date:	Tue, 16 May 2023	Prob (F-statistic):	0.00
Time:	16:55:28	Log-Likelihood:	19352.
No. Observations:	9504	AIC:	-3.868e+04
Df Residuals:	9491	BIC:	-3.859e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0190	0.004	5.360	0.000	0.012	0.026
num_refill_req_l3m	-0.0001	0.001	-0.115	0.908	-0.002	0.002
Trans_issue_l1y	-0.0151	0.001	-19.975	0.000	-0.017	-0.014
Competitor_in_mkt	0.0011	0.003	0.398	0.691	-0.004	0.007
retail_shop_num	0.0032	0.002	1.315	0.189	-0.002	0.008
distributor_num	0.0019	0.001	1.358	0.175	-0.001	0.005
dist_from_hub	0.0005	0.001	0.354	0.723	-0.002	0.003
WH_workers_num	0.0031	0.002	1.374	0.169	-0.001	0.008
WH_storage_iss_reported_l3m	0.8890	0.002	583.688	0.000	0.886	0.892
wh_breakdown_l3m	-0.0238	0.001	-19.089	0.000	-0.026	-0.021
govt_check_l3m	-0.0018	0.001	-1.486	0.137	-0.004	0.001
WH_Temp_reg_1	0.0177	0.001	24.062	0.000	0.016	0.019
Zone 6	-0.0016	0.001	-2.280	0.023	-0.003	-0.000

Fig29: After dropping dummy+Summary_Cluster1

In stats summary we see, there are variables whose p values are greater than 0.05. These variables are not significant; we are dropping those variables one by one.

We have dropped all the variables that are not significant and developed final model as below



OLS Regression Results

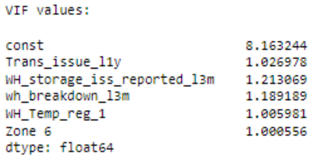
Dep. Variable:	prod_wg_ton	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	8.207e+04
Date:	Tue, 16 May 2023	Prob (F-statistic):	0.00
Time:	16:58:21	Log-Likelihood:	19348.
No. Observations:	9504	AIC:	-3.868e+04
Df Residuals:	9498	BIC:	-3.864e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0243	0.001	26.235	0.000	0.022	0.026
Trans_issue_l1y	-0.0152	0.001	-20.030	0.000	-0.017	-0.014
WH_storage_iss_reported_l3m	0.8890	0.002	583.829	0.000	0.886	0.892
wh_breakdown_l3m	-0.0238	0.001	-19.074	0.000	-0.026	-0.021
WH_Temp_reg_1	0.0176	0.001	24.899	0.000	0.016	0.019
Zone 6	-0.0014	0.001	-2.074	0.038	-0.003	-7.85e-05

Omnibus:	2550.574	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11595.880
Skew:	1.238	Prob(JB):	0.00
Kurtosis:	7.812	Cond. No.	7.19

Fig30: Summary-After dropping non-Significant-Cluster1

VIF:



```
VIF values:

const                          8.163244
Trans_issue_l1y                1.026978
WH_storage_iss_reported_l3m    1.213069
wh_breakdown_l3m               1.189189
WH_Temp_reg_1                  1.005981
Zone 6                         1.000556
dtype: float64
```

Fig31: VIF-After dropping non-Significant-Cluster1

Linearity Test:

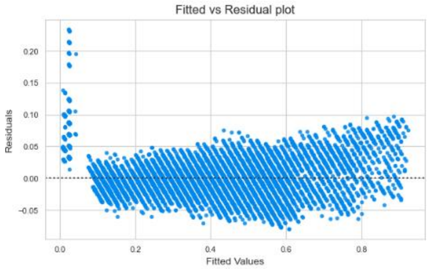


Fig32: Fitted Vs Residual plot-cluster1

We can see that data is randomly distributed.

Normality Test:

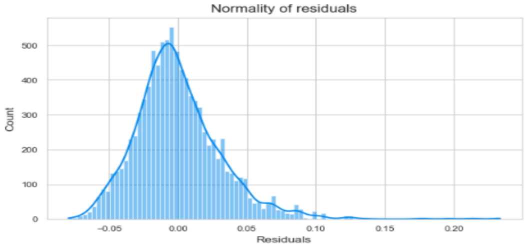


Fig33: Normality of Residuals-cluster1

The residual terms are normally distributed.

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

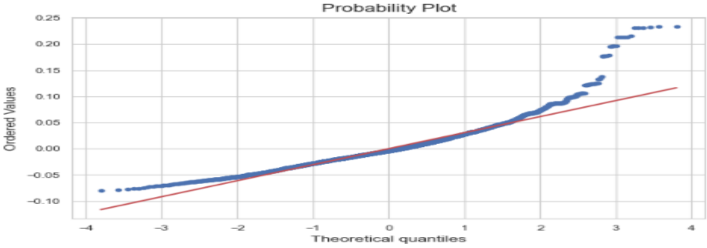


Fig34: Q-Q plot-cluster1

Most of the points lie on the straight line.

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

* Null hypothesis - Data is normally distributed.

* Alternate hypothesis - Data is not normally distributed.

Shapiro test resulted in p value less than 0.05 , the residuals are not normally distributed as per shapiro test.

ShapiroResult(statistic=0.9408862590789795, pvalue=0.0)

TEST FOR HOMOSCEDASTICITY:

The null and alternate hypotheses of the goldfeldquandt test are as follows:

Null hypothesis : Residuals are homoscedastic

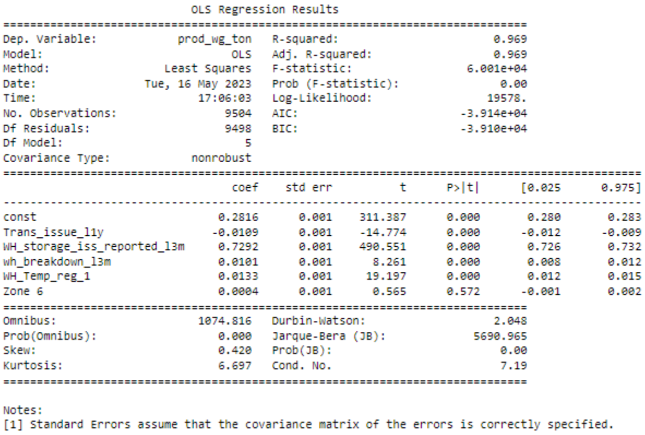
Alternate hypothesis : Residuals have hetroscedasticity

Result of test: [('F statistic', 1.0730621264796454), ('p-value', 0.007578631289403552)]

p-value =0.007 < 0.05 , we can say that the residuals are not homoscedastic.

So square root transforming is performed on target variable to make residuals to homoscedastic.

Final Model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            prod_wg_ton   R-squared:                       0.969
Model:                            OLS   Adj. R-squared:                  0.969
Method:                 Least Squares   F-statistic:                 6.001e+04
Date:                Tue, 16 May 2023   Prob (F-statistic):               0.00
Time:                        17:06:03   Log-Likelihood:                 19578.
No. Observations:                9504   AIC:                        -3.914e+04
Df Residuals:                    9498   BIC:                        -3.910e+04
Df Model:                           5
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                           0.2816      0.001    311.387      0.000       0.280       0.283
Trans_issue_l1y                -0.0109      0.001    -14.774      0.000      -0.012      -0.009
WH_storage_iss_reported_l3m     0.7292      0.001    490.551      0.000       0.726       0.732
wh_breakdown_l3m                0.0101      0.001      8.261      0.000       0.008       0.012
WH_Temp_reg_1                   0.0133      0.001     19.197      0.000       0.012       0.015
Zone 6                          0.0004      0.001      0.565      0.572      -0.001       0.002
==============================================================================
Omnibus:                     1074.816   Durbin-Watson:                   2.048
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5690.965
Skew:                           0.420   Prob(JB):                         0.00
Kurtosis:                       6.697   Cond. No.                         7.19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig35: Final Model-Cluster1

Observations:

R-squared of the model is 0.969 and adjusted R-squared is 0.969, which shows that the model is able to explain ~96% variance in the data. This is quite good. A unit increase in the warehouse storage issue reported in last 3 months will result in a 0.729 unit increase in the product weight, all other variables remaining constant. A unit decrease in the Transportation issues in last one year will result in a 0.0109 unit decrease in the product weight, all other variables remaining constant.

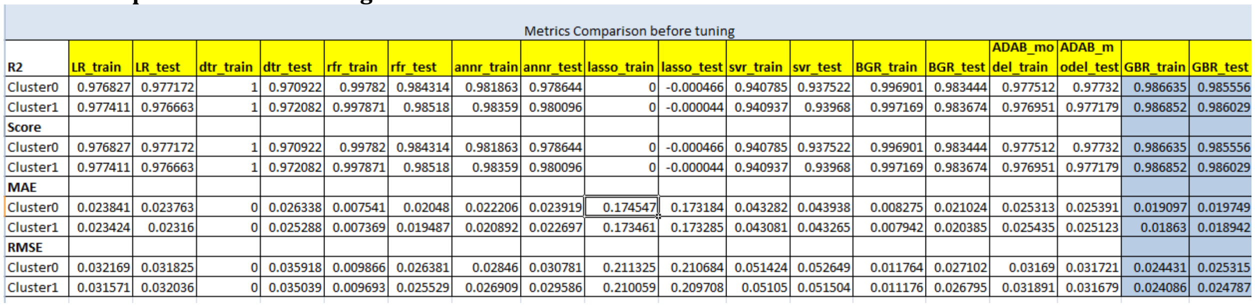
As per linear regression-cluster0, warehouse storage issue has the high importance while the Zone6 has least importance among the important variables.

Model:

product weight = 0.2815546357201927 + -0.010921538778672753 * (Trans_issue_l1y) + 0.7292044800765384 * (WH_storage_iss_reported_l3m) + 0.010069445133950456 * (wh_breakdown_l3m) + 0.013269385796542 * (WH_Temp_reg_1) + 0.0003809147538977181 * (Zone 6)

Developing different models on the Maximum Absolute Scaled data as this gave best results in linear regression model.

Models Comparison before tuning:



Metrics Comparison before tuning

R2	LR_train	LR_test	dtr_train	dtr_test	rfr_train	rfr_test	annr_train	annr_test	lasso_train	lasso_test	svr_train	svr_test	BGR_train	BGR_test	ADAB_model_train	ADAB_model_test	GBR_train	GBR_test
Cluster0	0.976827	0.977172	1	0.970922	0.99782	0.984314	0.981863	0.978644	0	-0.000466	0.940785	0.937522	0.996901	0.983444	0.977512	0.97732	0.986635	0.985556
Cluster1	0.977411	0.976663	1	0.972082	0.997871	0.98518	0.98359	0.980096	0	-0.000044	0.940937	0.93968	0.997169	0.983674	0.976951	0.977179	0.986852	0.986029
Score																		
Cluster0	0.976827	0.977172	1	0.970922	0.99782	0.984314	0.981863	0.978644	0	-0.000466	0.940785	0.937522	0.996901	0.983444	0.977512	0.97732	0.986635	0.985556
Cluster1	0.977411	0.976663	1	0.972082	0.997871	0.98518	0.98359	0.980096	0	-0.000044	0.940937	0.93968	0.997169	0.983674	0.976951	0.977179	0.986852	0.986029
MAE																		
Cluster0	0.023841	0.023763	0	0.026338	0.007541	0.02048	0.022206	0.023919	0.174547	0.173184	0.043282	0.043938	0.008275	0.021024	0.025313	0.025391	0.019097	0.019749
Cluster1	0.023424	0.02316	0	0.025288	0.007369	0.019487	0.020892	0.022697	0.173461	0.173285	0.043081	0.043265	0.007942	0.020385	0.025435	0.025123	0.01863	0.018942
RMSE																		
Cluster0	0.032169	0.031825	0	0.035918	0.009866	0.026381	0.02846	0.030781	0.211325	0.210684	0.051424	0.052649	0.011764	0.027102	0.03169	0.031721	0.024431	0.025315
Cluster1	0.031571	0.032036	0	0.035039	0.009693	0.025529	0.026909	0.029586	0.210059	0.209708	0.05105	0.051504	0.011176	0.026795	0.031891	0.031679	0.024086	0.024787

Fig36: Models Comparison-before tuning

Inferences:

- Among all models developed including Ensembling techniques, we see Gradient boosting regressor gave the best results with 98% aacuracy on both clusters. MAE and RMSE for on both the train and test are very low compared to other models.
- Decision tree regressor is slightly overfitting in case of both the clusters.
- Lasso regressor gave least r2 value.

Feature importance apart from warehouse storage issue variable in Gradient boosting Regressor after removing warehouse storage issue variable as it is occupying highest importance.

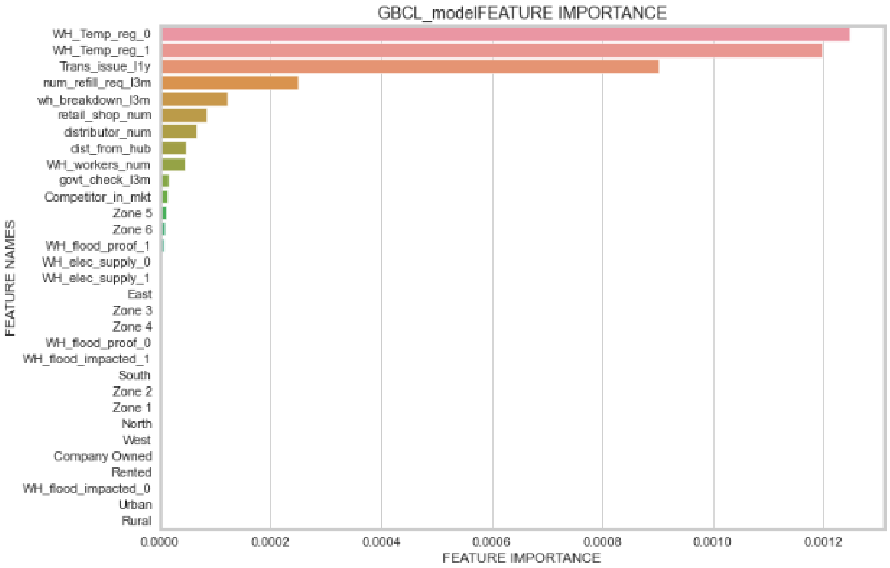
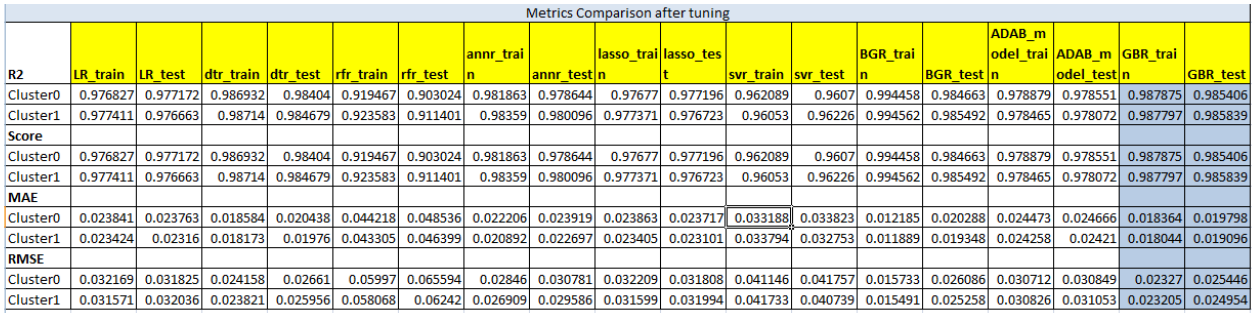


Fig37: Feature importance as per gradient boosting

After tuning:

Tuning Parameter:

Neural Network Regressor: hidden_layer_sizes=(500),random_state=123, max_iter=10000,activation= relu,solver= 'adam'
Random Forest Regressor: random_state=123,max_depth= 10,max_features= 6,min_samples_leaf= 3, min_samples_split= 30, n_estimators=300
Decision Tree Regressor: random_state=123,max_depth= 15, min_samples_leaf= 30, min_samples_split=15
Lasso Regressor: alpha=0.0001
Support Vector Regression: C= 1000, gamma= 0.01, kernel= 'rbf'
BaggingRegressor: random_state=123,max_samples=0.5,n_estimators=300
AdaBoostRegressor: random_state=123,loss=' exponential ',n_estimators=10,learning_rate=1.2
GradientBoostingRegressor:random_state=123,loss='squared_error',max_depth=4,criterion='friedman_mse', subsample=0.7,learning_rate = 0.1,max_features= 'auto',n_estimators=100



Metrics Comparison after tuning

R2	LR_train	LR_test	dtr_train	dtr_test	rfr_train	rfr_test	annr_train	annr_test	lasso_train	lasso_test	svr_train	svr_test	BGR_train	BGR_test	ADAB_model_train	ADAB_model_test	GBR_train	GBR_test
Cluster0	0.976827	0.977172	0.986932	0.98404	0.919467	0.903024	0.981863	0.978644	0.97677	0.977196	0.962089	0.9607	0.994458	0.984663	0.978879	0.978551	0.987875	0.985406
Cluster1	0.977411	0.976663	0.98714	0.984679	0.923583	0.911401	0.98359	0.980096	0.977371	0.976723	0.96053	0.96226	0.994562	0.985492	0.978465	0.978072	0.987797	0.985839
Score																		
Cluster0	0.976827	0.977172	0.986932	0.98404	0.919467	0.903024	0.981863	0.978644	0.97677	0.977196	0.962089	0.9607	0.994458	0.984663	0.978879	0.978551	0.987875	0.985406
Cluster1	0.977411	0.976663	0.98714	0.984679	0.923583	0.911401	0.98359	0.980096	0.977371	0.976723	0.96053	0.96226	0.994562	0.985492	0.978465	0.978072	0.987797	0.985839
MAE																		
Cluster0	0.023841	0.023763	0.018584	0.020438	0.044218	0.048536	0.022206	0.023919	0.023863	0.023717	0.033188	0.033823	0.012185	0.020288	0.024473	0.024666	0.018364	0.019798
Cluster1	0.023424	0.02316	0.018173	0.01976	0.043305	0.046399	0.020892	0.022697	0.023405	0.023101	0.033794	0.032753	0.011889	0.019348	0.024258	0.02421	0.018044	0.019096
RMSE																		
Cluster0	0.032169	0.031825	0.024158	0.02661	0.05997	0.065594	0.02846	0.030781	0.032209	0.031808	0.041146	0.041757	0.015733	0.026086	0.030712	0.030849	0.02327	0.025446
Cluster1	0.031571	0.032036	0.023821	0.025956	0.058068	0.06242	0.026909	0.029586	0.031599	0.031994	0.041733	0.040739	0.015491	0.025258	0.030826	0.031053	0.023205	0.024954

Fig38: Model comparison after tuning

Inferences:

- ➢ Gradient boosting regression gave the best results even after tuning the parameters.
- ➢ Gradient boosting regressor gave 98.7 % accuracy on train and 98.5 accuracy on test.
- ➢ We don't see any overfit or underfit.
- ➢ MAE and RMSE values thru the model Gradient boosting regressor is the least.
- ➢ Below is the coefficient as per the gradient boosting regressor:

```
array([3.58050249e-04, 1.04752197e-03, 1.15192512e-04, 3.65229960e-04,
       2.21179606e-04, 2.97270031e-04, 1.68429763e-04, 9.94289984e-01,
       2.45183993e-04, 1.69653640e-04, 4.72296767e-06, 1.94938846e-06,
       1.26162075e-03, 1.22277045e-03, 1.05089285e-05, 1.92940987e-05,
       2.47014601e-06, 1.26864140e-05, 1.31594741e-05, 2.80680380e-05,
       1.77325881e-05, 1.64723139e-05, 2.79941222e-05, 3.94647633e-06,
       0.00000000e+00, 0.00000000e+00, 9.98063723e-06, 0.00000000e+00,
       8.51528890e-06, 1.37168740e-05, 1.82844343e-05, 2.84112758e-05])
```

Apart from warehouse storage issue, weightage od othervariables as per the gradient boosting regressor:

For cluster1:

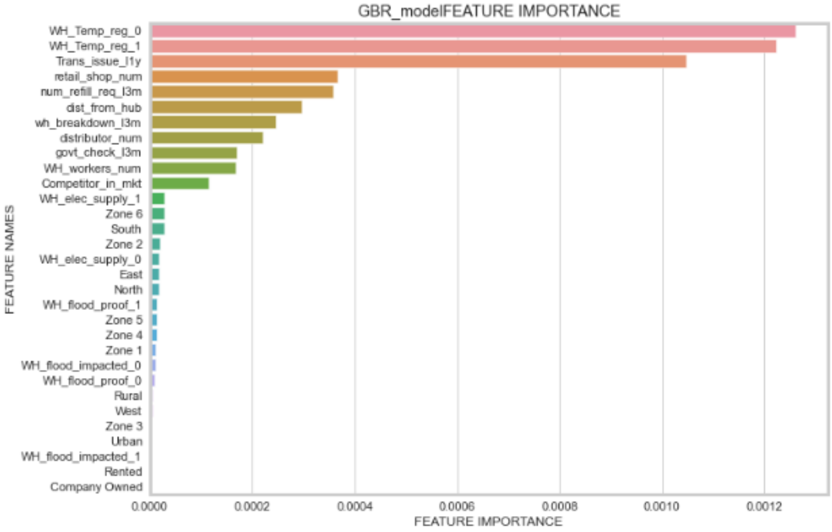


Fig39: Feature imp-plot-GBR-1

For Cluster0:

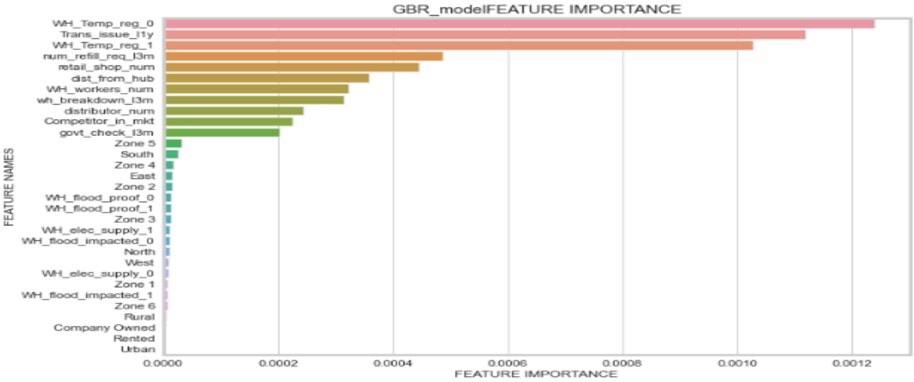


Fig40: Feature imp-plot-GBR-0

Business Implications:

For both the clusters, warehouse storage issue in the last 1 year –variable has highest weight.

As the number of storage issues increases, it increases the product weight as well which means when there is some storage issue, there will not be smooth delivery of products and it is not flexible to track and manage inventory at the particular place. Inorder to fulfill the demands in the area where the warehouse is established, more and more products need to be sent to the warehouse if there is storage issue.

If storage issue is minimal, then it reduces the transportation costs, increases the flexibility of delivering the products and reduces the workers need.

It also helps for proper managing of inventory.

Temperature regulator is second most important variable in both the clusters. High temperatures can also lead to spoilage of products, especially those sensitive to environmental conditions such as pharmaceuticals, or foodstuffs. When the quality and purity of the products rely on correct storage it is critical that a warehouse temperature and humidity monitoring system is employed.

Transportation Issue in the last one year has next weight in the Cluster1 and 2nd importance in cluster0. If transportation issue increases, delivery of products on time is not possible.

Maintenance and insurance cost, fuel cost increases which impacts the business.

Retail shop number is directly proportional to the product weight.

In both clusters, Warehouse refill is important factor as where they must be re-ordered to avoid stock shortages.

In Cluster1, distance from hub, workers count, distributors count,competitors, warehouse breakage, power back up and government check variables also has some amount of impact.

In Cluster0, distance from hub, workers count, distributors count, competitors, warehouse breakage, Zone5,North Zoneand government check variables also has some amount of impact.