CAPSTONE PROJECT REPORT

Submitted By

Vinit Sharma

1. Introduction:

Problem Statements:

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.

Need of the study/project

By utilizing the last 2 years of dataset, we need to predict the number of products that needs to ship from each warehouse to the stores.

Understanding business/social opportunity

Based on the dataset variables, we have to analysis the business opportunities. For examples, we have two type of warehouse details like company owned and rented. We can be look for the benefits prospective which can be monetary based on demand supply requirements.

2. EDA and Business Implication

Entire dataset collection is done based on last 2 year of history of the supply chain management of instant noodles. Dataset has 25000 rows and 21 variables columns. We have showcased the descriptive details of the numeric variables. Details include the total count, mean, min, quartile and max information.

	count	mean	std	min	25%	50%	75%	max
num_refill_req_l3m	25000.00	4.09	2.61	0.00	2.00	4.00	6.00	8.00
transport_issue_l1y	25000.00	0.77	1.20	0.00	0.00	0.00	1.00	5.00
Competitor_in_mkt	25000.00	3.10	1.14	0.00	2.00	3.00	4.00	12.00
retail_shop_num	25000.00	4985.71	1052.83	1821.00	4313.00	4859.00	5500.00	11008.00
distributor_num	25000.00	42.42	16.06	15.00	29.00	42.00	56.00	70.00
flood_impacted	25000.00	0.10	0.30	0.00	0.00	0.00	0.00	1.00
flood_proof	25000.00	0.05	0.23	0.00	0.00	0.00	0.00	1.00
electric_supply	25000.00	0.66	0.47	0.00	0.00	1.00	1.00	1.00
dist_from_hub	25000.00	163.54	62.72	55.00	109.00	164.00	218.00	271.00
workers_num	24010.00	28.94	7.87	10.00	24.00	28.00	33.00	98.00
storage_issue_reported_l3m	25000.00	17.13	9.16	0.00	10.00	18.00	24.00	39.00
temp_reg_mach	25000.00	0.30	0.46	0.00	0.00	0.00	1.00	1.00
wh_breakdown_l3m	25000.00	3.48	1.69	0.00	2.00	3.00	5.00	6.00
govt_check_l3m	25000.00	18.81	8.63	1.00	11.00	21.00	26.00	32.00
product_wg_ton	25000.00	22102.63	11607.76	2065.00	13059.00	22101.00	30103.00	55151.00

FIGURE 1: DESCRIPTIVE ANALYSIS

Dataset variables include float, int and object datatype. It has a 6 object, 14 int and 1 float datatype variable. Based on information, we can observe that some variables have missing values such as workers_sum and approved_wh_govt_certificate. This dataset has total 21 columns and 25000 entries.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 21 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Location_type                 25000 non-null  object
 1   WH_capacity_size              25000 non-null  object
 2   zone                          25000 non-null  object
 3   WH_regional_zone              25000 non-null  object
 4   num_refill_req_l3m            25000 non-null  int64
 5   transport_issue_l1y           25000 non-null  int64
 6   Competitor_in_mkt             25000 non-null  int64
 7   retail_shop_num               25000 non-null  int64
 8   wh_owner_type                 25000 non-null  object
 9   distributor_num               25000 non-null  int64
 10  flood_impacted                25000 non-null  int64
 11  flood_proof                   25000 non-null  int64
 12  electric_supply               25000 non-null  int64
 13  dist_from_hub                 25000 non-null  int64
 14  workers_num                   24010 non-null  float64
 15  storage_issue_reported_l3m    25000 non-null  int64
 16  temp_reg_mach                 25000 non-null  int64
 17  approved_wh_govt_certificate  24092 non-null  object
 18  wh_breakdown_l3m              25000 non-null  int64
 19  govt_check_l3m                25000 non-null  int64
 20  product_wg_ton                25000 non-null  int64
dtypes: float64(1), int64(14), object(6)
memory usage: 4.0+ MB
```

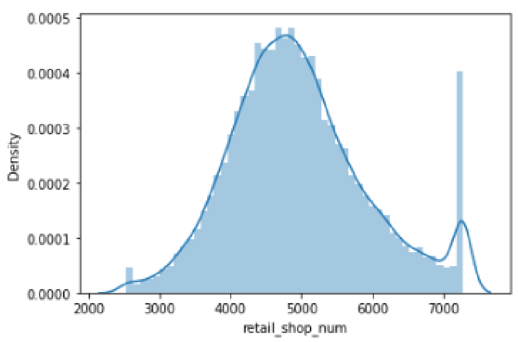
FIGURE 2: VARIABLE INFO

Exploratory Data Analysis

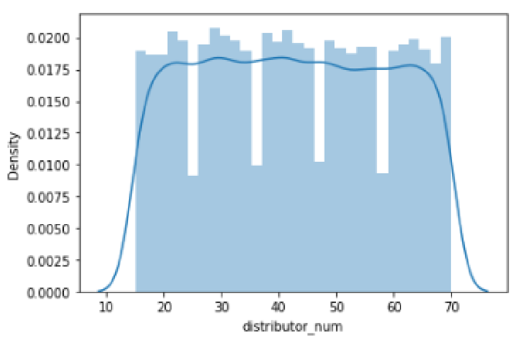
Univariate analysis:

Descriptive and distplot analysis has been showcased for Retail_shop_num, distributor_num , dist_from_hub, workers_num, product_wg_ton and govt_check_l3m variables. Retail_shop_num and worker num data are normally distributed. Govt_check_l3m and product_wg_ton are left skewed and right skewed respectively. Distributor_num , and dist_from_hub dataset is balanced distributed.

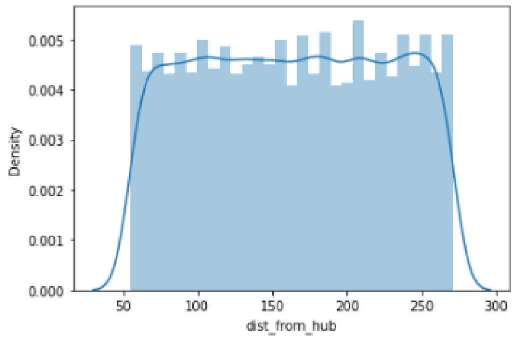
```
Description of retail_shop_num
----------------------------------------------------------
count    25000.00
mean      4958.89
std        969.40
min       2532.50
25%       4313.00
50%       4859.00
75%       5500.00
max       7280.50
Name: retail_shop_num, dtype: float64

----------------------------------------------------------
Distplot of retail_shop_num
```



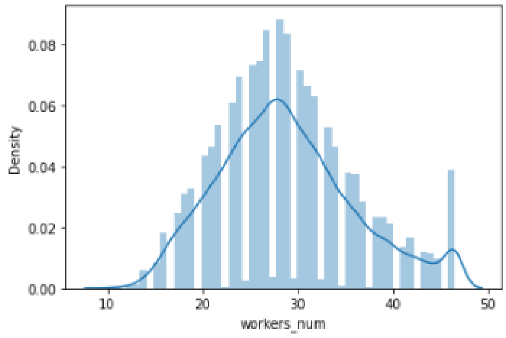

```
Description of distributor_num
----------------------------------------------------------
count    25000.00
mean        42.42
std         16.06
min         15.00
25%         29.00
50%         42.00
75%         56.00
max         70.00
Name: distributor_num, dtype: float64

----------------------------------------------------------
Distplot of distributor_num
```



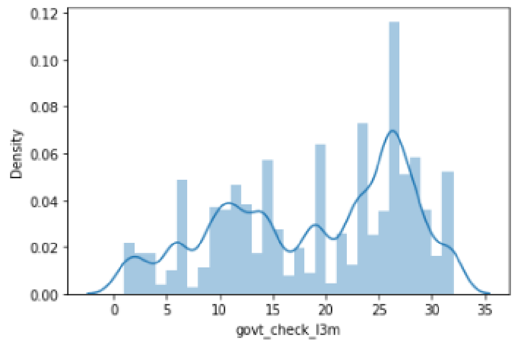
```
Description of dist_from_hub
------------------------------------------------------------
count    25000.00
mean       163.54
std         62.72
min         55.00
25%        109.00
50%        164.00
75%        218.00
max        271.00
Name: dist_from_hub, dtype: float64

------------------------------------------------------------
Distplot of dist_from_hub
```




```
Description of workers_num
------------------------------------------------------------
count    25000.00
mean        28.78
std          7.17
min         10.50
25%         24.00
50%         28.00
75%         33.00
max         46.50
Name: workers_num, dtype: float64

------------------------------------------------------------
Distplot of workers_num
```




```
Description of govt_check_l3m
------------------------------------------------------------
count    25000.00
mean        18.81
std          8.63
min          1.00
25%         11.00
50%         21.00
75%         26.00
max         32.00
Name: govt_check_l3m, dtype: float64

------------------------------------------------------------
Distplot of govt_check_l3m
```




```
Description of product_wg_ton
------------------------------------------------------------
count    25000.00
mean     22102.63
std      11607.76
min       2065.00
25%      13059.00
50%      22101.00
75%      30103.00
max      55151.00
Name: product_wg_ton, dtype: float64

------------------------------------------------------------
Distplot of product_wg_ton
```

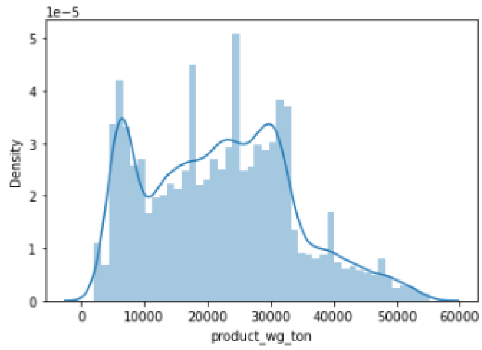
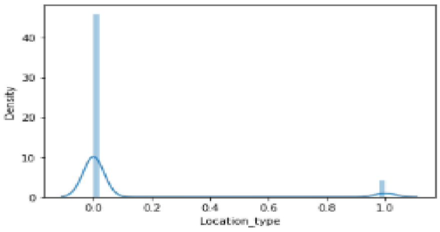


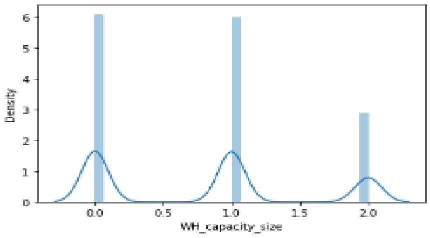
FIGURE 3: DESCRIPTIVE AND DISTPLOT ANALYSIS FOR CONTINUOUS VARIABLE

Based on categorical feature analysis, we have seen the density plots and their freq. values in the dataset. We have plot the distplot and done the categorical descriptive analysis for the variables as well.

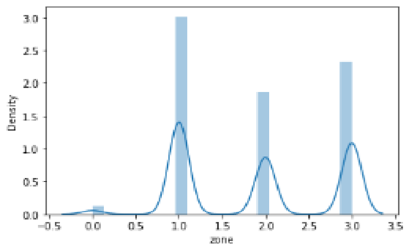
```
Description of Location_type
---------------------------------------------------------------
count    25000.00
mean         0.08
std          0.27
min          0.00
25%          0.00
50%          0.00
75%          0.00
max          1.00
Name: Location_type, dtype: float64

---------------------------------------------------------------
Distplot of Location_type
```



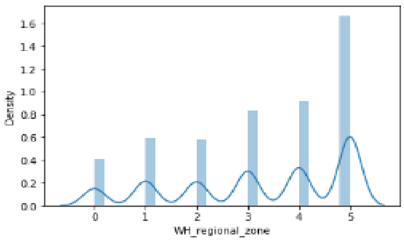
```
Description of WH_capacity_size
---------------------------------------------------------------
count    25000.00
mean         0.79
std          0.74
min          0.00
25%          0.00
50%          1.00
75%          1.00
max          2.00
Name: WH_capacity_size, dtype: float64

---------------------------------------------------------------
Distplot of WH_capacity_size
```



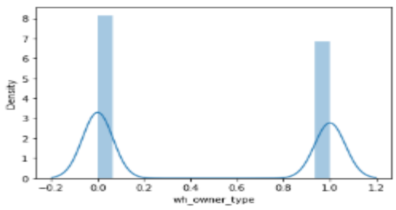
```
Description of zone
---------------------------------------------------------------
count    25000.00
mean         1.87
std          0.88
min          0.00
25%          1.00
50%          2.00
75%          3.00
max          3.00
Name: zone, dtype: float64

---------------------------------------------------------------
Distplot of zone
```



```
Description of WH_regional_zone
---------------------------------------------------------------
count    25000.00
mean         3.25
std          1.67
min          0.00
25%          2.00
50%          4.00
75%          5.00
max          5.00
Name: WH_regional_zone, dtype: float64

---------------------------------------------------------------
Distplot of WH_regional_zone
```



```
Description of wh_owner_type
---------------------------------------------------------------
count    25000.00
mean         0.46
std          0.50
min          0.00
25%          0.00
50%          0.00
75%          1.00
max          1.00
Name: wh_owner_type, dtype: float64

---------------------------------------------------------------
Distplot of wh_owner_type
```



```
Description of approved_wh_govt_certificate
---------------------------------------------------------------
count    25000.00
mean         2.17
std          1.45
min          0.00
25%          1.00
50%          2.00
75%          4.00
max          4.00
Name: approved_wh_govt_certificate, dtype: float64

---------------------------------------------------------------
Distplot of approved_wh_govt_certificate
```

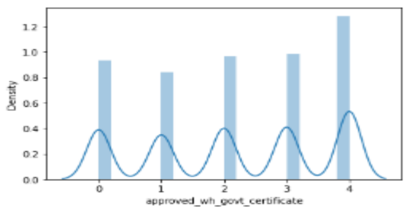


FIGURE 4: DESCRIPTIVE AND DISTPLOT ANALYSIS FOR CATEGORICAL VARIABLE

We have done the skewness anaylsis for the variables. Based on the results, flood_proof, flood_impacted and transport_issue_l1y is mostly positive skewed. Electric_supply and govt_check_l3m is negatively skewed. Rest other variables are mostly balanced.

```
num_refill_req_l3m            -0.08
transport_issue_l1y            1.61
Competitor_in_mkt              0.80
retail_shop_num                0.44
distributor_num                0.02
flood_impacted                 2.70
flood_proof                    3.92
electric_supply               -0.66
dist_from_hub                 -0.01
workers_num                    0.43
storage_issue_reported_l3m     0.11
temp_reg_mach                  0.86
wh_breakdown_l3m              -0.07
govt_check_l3m                -0.36
product_wg_ton                 0.33
dtype: float64
```

FIGURE 5: SKEWNESS ANALYSIS FOR VARIABLES

Based on plot, we can conclude that most of the data is rural inclined compared to urban set. It could be for multiple reasons such as demand, store availability, transportation convenience.

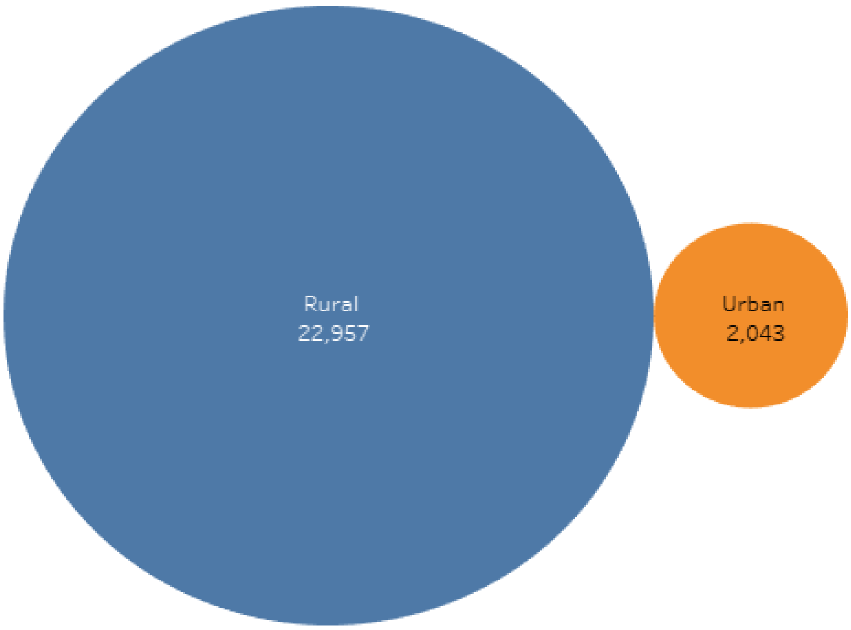


FIGURE 6: BUBBLE PLOTS FOR ANALYSIS

Bivariate Analysis:

Based on below analysis, we can showcased North zone with zone 6 have major product weight ton in the process. Bar plot complete the analysis on zone with regional zone and product requirements.

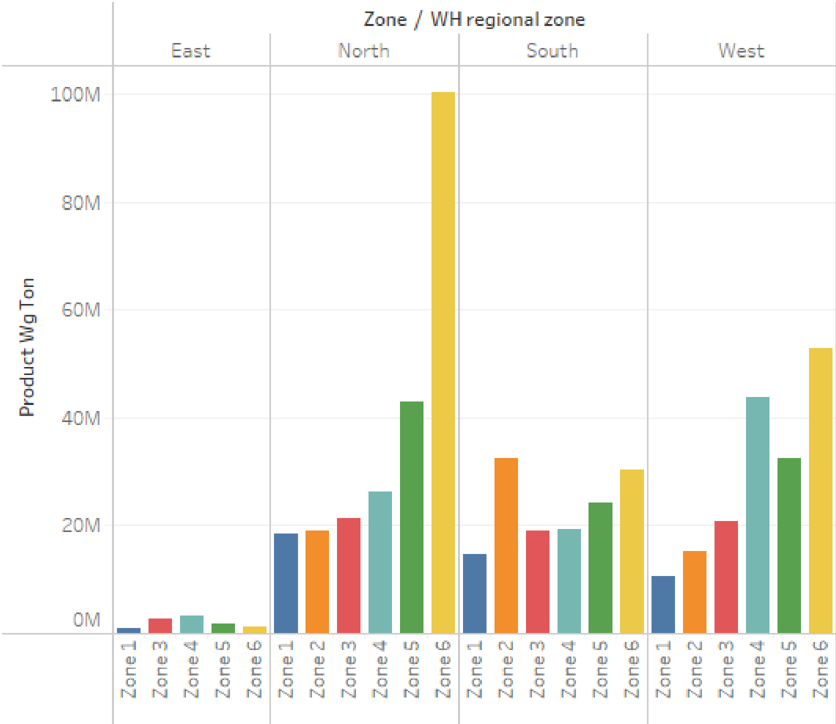
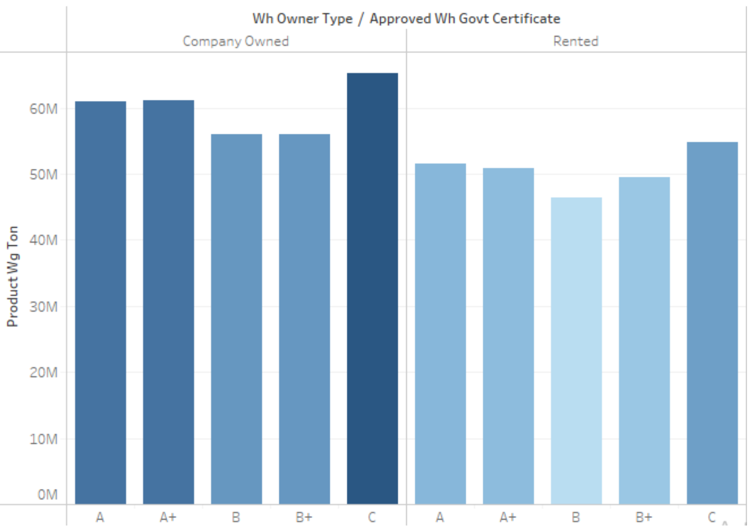
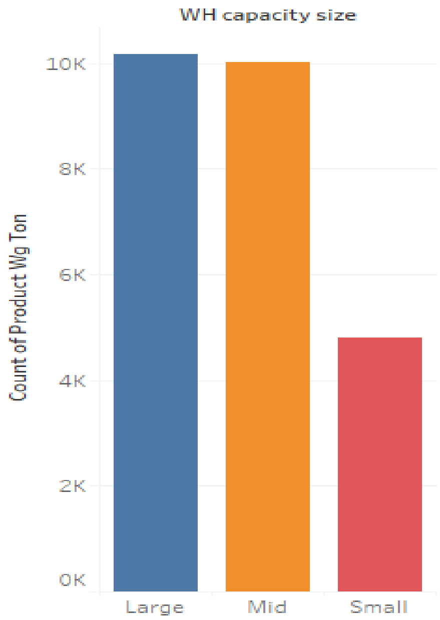


FIGURE 7: BAR PLOT FOR ZONE AND WH REGIONAL ZONE WITH PRODUCT INFO

This bar plot showcased warehouse owner type and approved warehouse govt certificate with respect to product wg ton requirement. As per plot, we have seen almost all have same frequency.



Wh capacity size with respect to product wg ton is plotted in a form of bar plot.



Multivariate Analysis:

Below plot shows the correlation metrics among the variables. We can figure that workers_sum with electric supply, storage_issue_reported_l3m with product wg ton are highly positively correlated. Rest variables are much neutral correlated.

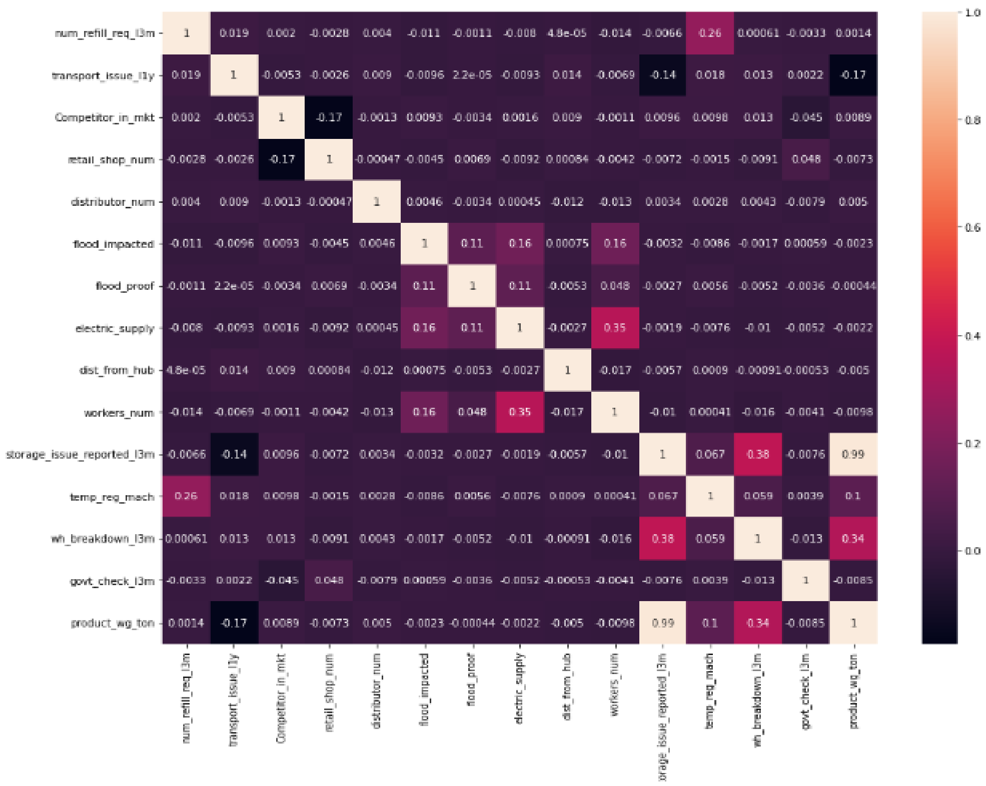


FIGURE 8: CORRELATION PLOT

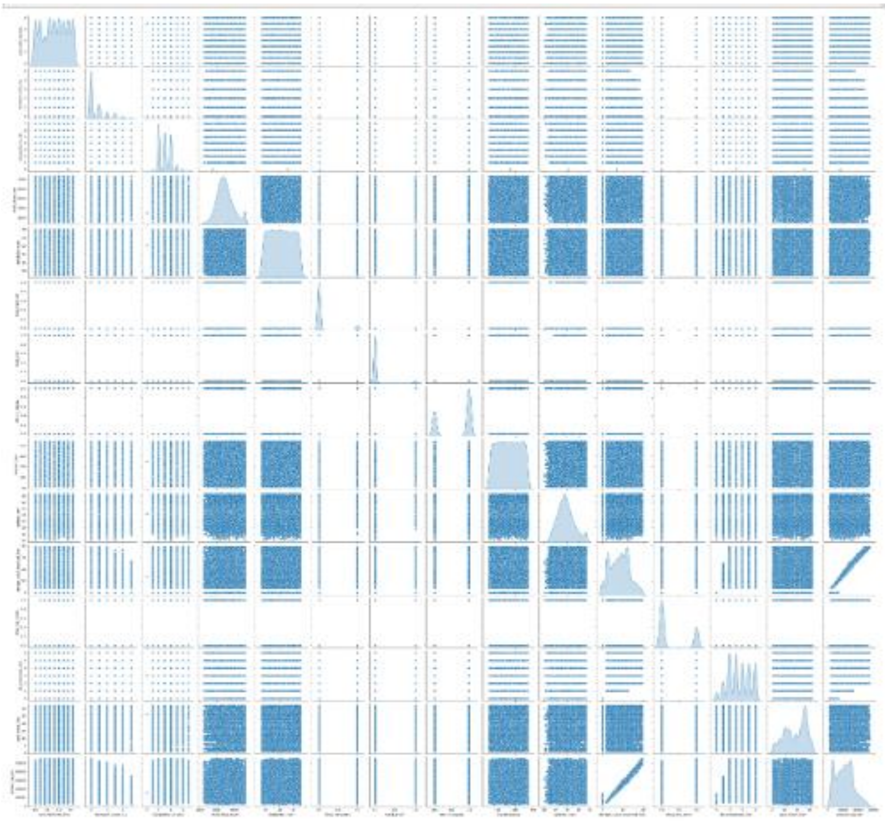


FIGURE 9: PAIR PLOT

3. Data Cleaning and Pre-processing:

Removal of unwanted variables (if applicable)

We have removed the 3 variables which are 'Ware_house_ID', 'WH_Manager_ID' and 'wh_est_year' from the dataset. “Ware_house_ID” and “WH_Manager_ID” has a unique data for each row which will not be useful for the model development. In a similar way, “wh_est_year” has a information related to year of warehouse establishment which has unique dates and multiple missing value. We can drop this variable as well from analysis.

Missing Value treatment (if applicable)

We have two variables with missing values. Imputation methods have been used to deal with such variable. To deal with workers_num, we have used KNN imputation method. Workers_sum includes numerical datasets. Approved_wh_govt_certified variables has categorical dataset so we have used” most_frequent” strategy to treat the missing value.

Outlier treatment (if required)

Based on dataset, we have figured out presence of outlier in 4 variables which are as follow:

```
Outlier value in transport_issue_l1y:  2943
Outlier value in Competitor_in_mkt:  96
Outlier value in retail_shop_num:  1174
Outlier value in workers_num:  346
```

Quartile method has been used to treat the outlier presence in Competitor_in_mkt, retail_shop_num and workers_num. All the three are contain very minimal numbers and also more of independent in natures. Variable such as workers_num is mostly depends on size of warehouse and kind of work. But its range ca n be bound to upper quartile. Transport_issue_l1y variable has the greatest number of outliers. It could b e due to multiple reasons such as distance from warehouse, frequency, delivery of quantity on retails, ex trinsic factor because of which we are not treating this variable for outlier treatment.

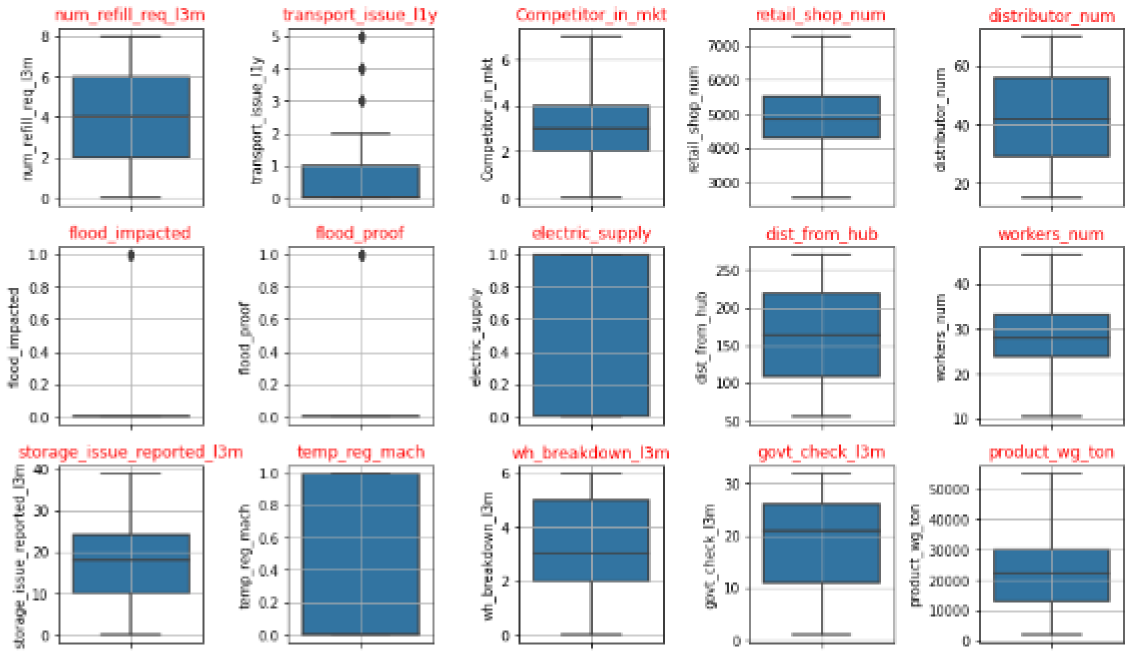


FIGURE 10: OUTLIER TREATED VARIABLES

Variable transformation (if applicable)

Dataset contains some variables in catego rical forms such as Location_type, WH_cap acity_size, zone, WH_regional_zone, wh_o wner_type and approved_wh_govt_certific ate. Label encoding technique has been us ed to convert the categorical variable to la bel forms. All these variables need to be co nverted into label form. This process will b e used for better predicting model building process.

```
feature: Location_type
[1 0]


feature: WH_capacity_size
[2 0 1]


feature: zone
[3 1 2 0]


feature: WH_regional_zone
[5 4 1 2 0 3]


feature: wh_owner_type
[1 0]


feature: approved_wh_govt_certificate
[0 1 4 2 3]
```

Based on dataset prospective, we have checked categorical variables to look for the balance in the dataset. Location_TYPE shows imbalance of the dataset. WH_CAPACITY_SIZE and ZONE is almost balance. Rest dataset values are much balanced. For location_type, we can conclude multiple store are available in rural areas and demand of instant noodles is more in rural area. So we are not look for treating the dataset.

```
LOCATION_TYPE : 2
Rural     22957
Urban      2043
Name: Location_type, dtype: int64

**************************************************

WH_CAPACITY_SIZE : 3
Large    10169
Mid      10020
Small     4811
Name: WH_capacity_size, dtype: int64

**************************************************

ZONE : 4
North    10278
West      7931
South     6362
East       429
Name: zone, dtype: int64

**************************************************
```

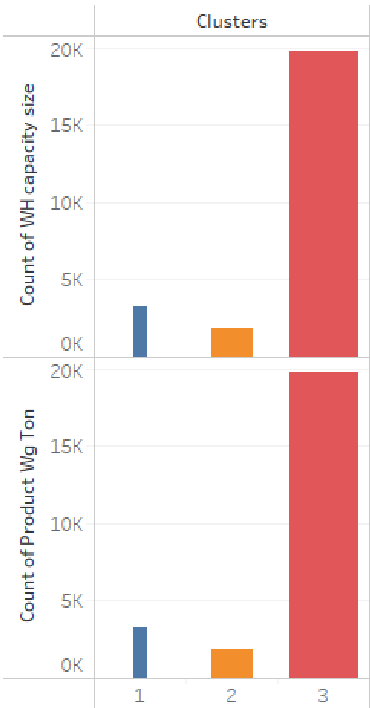
```
WH_REGIONAL_ZONE : 6
Zone 6    8339
Zone 5    4587
Zone 4    4176
Zone 2    2963
Zone 3    2881
Zone 1    2054
Name: WH_regional_zone, dtype: int64

**************************************************

WH_OWNER_TYPE : 2
Company Owned    13578
Rented           11422
Name: wh_owner_type, dtype: int64

**************************************************

APPROVED_WH_GOVT_CERTIFICATE : 5
C     6409
B+    4917
B     4812
A     4671
A+    4191
Name: approved_wh_govt_certificate, dtype: int64

**************************************************
```

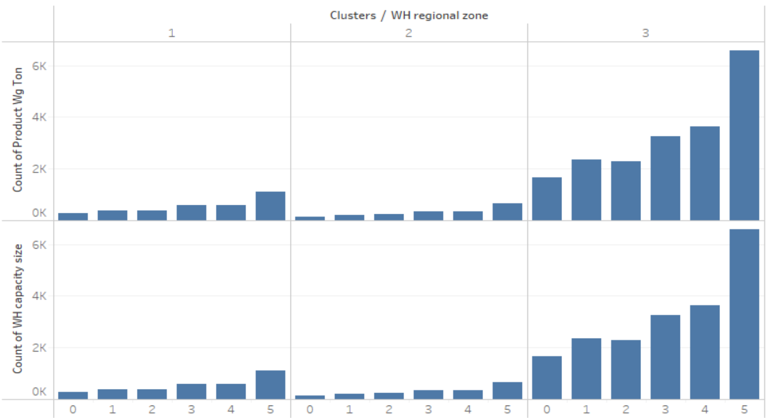
f)Addition of new variables (if required)

In our initial analysis, there is no addition of variable in the context. In later sense, we have added one clusters column for better analysis in study.

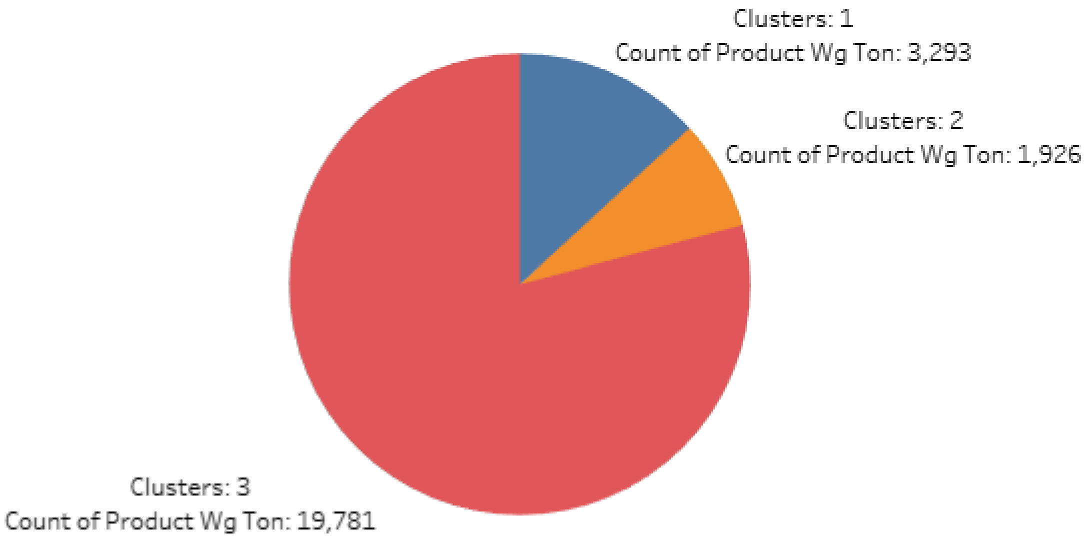
Business insights from EDA

As per the clustering approach on dataset, we have observed that major dataset is aligned with cluster 3. In each segment, we have to more focused towards the zone 4 to 6 in sense of demand frequency. In such segment, we have to prepare for more transportation frequency issues, extrinsic disturbance.





Cluster 3 is covered the major portion in the segments as seen in pie chart.



4. Model building & Validations

Based on problem statement, we have built multiple models such as Linear, decision tree and random forest for the predictive purpose.

Linear Regression:

```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)
```

Decision Tree Regression:

```
dt_model = DecisionTreeRegressor()
dt_model.fit(X_train, y_train)
```

Random Forest Regression:

```
rf_model = RandomForestRegressor(n_estimators = 100, random_state = 0)
rf_model.fit(X_train, y_train)
```

Bagging Regression:

```
bgrg = BaggingRegressor(n_estimators=50,random_state=1,n_jobs=-1)

bgrg = bgrg.fit(X_train, y_train)
```

Test of predictive model against the test set using various appropriate performance metrics

Linear Regression:

Below is the result for the linear regression model, we have intercept and coefficient values for all the variables. Coefficient value are in both positive and negative in nature. Some variables such as storage issue reported l3m and temp reg mach is highly positive which signifies a value change in storage issue reported l3m will effect the product_wgh_ton in positive side. In a same way, some negative coefficient variables such as location_type, transport_issue_l1y , wh_breakdown_l1m will negative effect the target variables.

```
The intercept for our model is 1684.3273458754556
```

```
The coefficient for Location_type is -109.00430606114132
The coefficient for WH_capacity_size is 11.128120809293875
The coefficient for zone is -3.8314078256313655
The coefficient for WH_regional_zone is -7.0222232835753955
The coefficient for num_refill_req_l3m is -2.5179083996127423
The coefficient for transport_issue_l1y is -310.6798423163092
The coefficient for Competitor_in_mkt is -8.001649806251871
The coefficient for retail_shop_num is -0.01424719652025266
The coefficient for wh_owner_type is 14.018823281343089
The coefficient for distributor_num is 1.1794545883088545
The coefficient for flood_impacted is 20.946341267744547
The coefficient for flood_proof is 54.6979951975392
The coefficient for electric_supply is 10.80469670504967
The coefficient for dist_from_hub is 0.2635033595973085
The coefficient for workers_num is -0.1651965163864737
The coefficient for storage_issue_reported_l3m is 1255.4020652578736
The coefficient for temp_reg_mach is 902.261707193937
The coefficient for approved_wh_govt_certificate is -105.06041634973246
The coefficient for wh_breakdown_l3m is -244.93902812645402
The coefficient for govt_check_l3m is -0.22604309621503305
```

These coefficient result signifies about the how much effect will be on the target variables based on the coefficient values change.

Adj R Square and RMSE values have been calculated and show the significance of the model. R square value for test shows 97.79 %, a model can predict the value of the target variable. We have other error also for the models.

Training Model Performance:

```
Mean Absolute Error (MAE): 1294.4932657882464
Mean Squared Error (MSE): 3139379.0247578584
Root Mean Squared Error (RMSE): 1771.8292877017973
Mean Absolute Percentage Error (MAPE): 0.09011731113672405
R^2: 0.9768486401787072
Adj R^2: 0.976830103520057
```

Test Model Performance:

```
Mean Absolute Error (MAE): 1275.0381190320231
Mean Squared Error (MSE): 2929496.7203814713
Root Mean Squared Error (RMSE): 1711.5772610026902
Mean Absolute Percentage Error (MAPE): 0.08844699146178421
R^2: 0.9779246658409846
Adj R^2: 0.9779069907265613
```

States Model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           product_wg_ton   R-squared:                       0.977
Model:                              OLS   Adj. R-squared:                  0.977
Method:                   Least Squares   F-statistic:                 3.688e+04
Date:                  Sun, 09 Oct 2022   Prob (F-statistic):               0.00
Time:                          13:41:30   Log-Likelihood:            -1.5573e+05
No. Observations:                 17500   AIC:                         3.115e+05
Df Residuals:                     17479   BIC:                         3.117e+05
Df Model:                            20
Covariance Type:              nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                    1684.3273    140.833     11.960      0.000    1408.280    1960.375
Location_type                -109.0043     49.267     -2.213      0.027    -205.572     -12.437
WH_capacity_size               11.1281     21.335      0.522      0.602     -30.691      52.947
zone                           -3.8314     15.706     -0.244      0.807     -34.616      26.954
WH_regional_zone               -7.0222      9.511     -0.738      0.460     -25.665      11.621
num_refill_req_l3m             -2.5179      5.333     -0.472      0.637     -12.971       7.935
transport_issue_l1y          -310.6798     11.319    -27.449      0.000    -332.865    -288.494
Competitor_in_mkt              -8.0016     12.334     -0.649      0.517     -32.177      16.174
retail_shop_num                -0.0142      0.014     -1.007      0.314      -0.042       0.013
wh_owner_type                  14.0188     27.951      0.502      0.616     -40.767      68.805
distributor_num                 1.1795      0.835      1.413      0.158      -0.456       2.815
flood_impacted                 20.9463     46.694      0.449      0.654     -70.579     112.471
flood_proof                    54.6980     60.242      0.908      0.364     -63.383     172.779
electric_supply                10.8047     31.020      0.348      0.728     -49.998      71.607
dist_from_hub                   0.2635      0.214      1.232      0.218      -0.156       0.683
workers_num                    -0.1652      2.030     -0.081      0.935      -4.145       3.815
storage_issue_reported_l3m   1255.4021      1.617    776.183      0.000    1252.232    1258.572
temp_reg_mach                 902.2617     30.546     29.538      0.000     842.389     962.134
approved_wh_govt_certificate -105.0604      9.556    -10.994      0.000    -123.791     -86.329
wh_breakdown_l3m             -244.9390      8.626    -28.395      0.000    -261.847    -228.031
govt_check_l3m                 -0.2260      1.655     -0.137      0.891      -3.470       3.018
==============================================================================
Omnibus:                     5937.490   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            36565.884
Skew:                           1.495   Prob(JB):                         0.00
Kurtosis:                       9.420   Cond. No.                     5.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

VIF Analysis:

```
Location_type ---> 1.0973445203552459
WH_capacity_size ---> 2.8108329811221857
zone ---> 5.538050924233202
WH_regional_zone ---> 6.038444709730886
num_refill_req_l3m ---> 3.6500741554963376
transport_issue_l1y ---> 1.4484223262164364
Competitor_in_mkt ---> 8.2387631139269
retail_shop_num ---> 20.570291726593563
wh_owner_type ---> 1.9339293443243404
distributor_num ---> 7.438537592906299
flood_impacted ---> 1.167041788669339
```

Based on VIF Analysis, retail_shop_num plays a most significant role in the variables importance. In a next view, Competitor_in_mkt, distributor_num , WH_regional_zone and Zone are the important factor of variable.

Decision Tree Regression:

Adj R Square and RMSE values have been calculated and show the significance of the model. R square value for test shows 98.7 %, a model can predict the value of the target variable. We have other error also for the models.

Training Model Performance:

```
Mean Absolute Error (MAE): 0.0
Mean Squared Error (MSE): 0.0
Root Mean Squared Error (RMSE): 0.0
Mean Absolute Percentage Error (MAPE): 0.0
R^2: 1.0
Adj R^2: 1.0
```

Test Model Performance:

```
Mean Absolute Error (MAE): 853.6946666666666
Mean Squared Error (MSE): 1607766.9186666666
Root Mean Squared Error (RMSE): 1267.977491388024
Mean Absolute Percentage Error (MAPE): 0.054773845005957646
R^2: 0.9878846111236623
Adj R^2: 0.9878749106641753
```

Random Forest Regression:

Adj R Square and RMSE values have been calculated and show the significance of the model. R square value for test shows 99.3 %, a model can predict the value of the target variable. We have other error also for the models.

Training Model Performance:

```
Mean Absolute Error (MAE): 266.03101428571426
Mean Squared Error (MSE): 132219.14545330856
Root Mean Squared Error (RMSE): 363.61950642575346
Mean Absolute Percentage Error (MAPE): 0.017123237632071565
R^2: 0.9990249495242488
Adj R^2: 0.9990241688280834
```

Test Model Performance:

```
Mean Absolute Error (MAE): 704.1833
Mean Squared Error (MSE): 884354.7646653465
Root Mean Squared Error (RMSE): 940.4013848699642
Mean Absolute Percentage Error (MAPE): 0.04566724102422662
R^2: 0.9933359109743045
Adj R^2: 0.9933305752210512
```

Bagging Regression Result:

Adj R Square and RMSE values have been calculated and show the significance of the model. R square value for test shows 99.3 %, a model can predict the value of the target variable. We have other error also for the models.

Training Model Performance:

```
Mean Absolute Error (MAE): 269.3208365714287
Mean Squared Error (MSE): 138293.41068082285
Root Mean Squared Error (RMSE): 371.8782202291805
Mean Absolute Percentage Error (MAPE): 0.017338525906530823
R^2: 0.9989801548375216
Adj R^2: 0.9989793382754795
```

Test Model Performance:

```
Mean Absolute Error (MAE): 707.7308266666665
Mean Squared Error (MSE): 894225.4403786667
Root Mean Squared Error (RMSE): 945.6349403330372
Mean Absolute Percentage Error (MAPE): 0.04597357056789496
R^2: 0.9932615301213645
Adj R^2: 0.9932561348134029
```

Interpretation of the model(s):

Based on the model result, we have seen that Random Forest is better model compare to other build models based on R square and RMSE values. Decision tree is the second better model in the process based on metrics analysis. In the linear regression model, Coefficient value of some variables are highly positive and negative which shows the variable significance on the result. In a state model, we have found many variables p values are more than 0.05 which can be neglect to enhance the model performance.

Hyperparameter Tuning for RF:

We have used **Grid Search technique** to tuning the model parameters. This tuning is done on random forest model. After completing the hyperparameter tuning process, we have found the best parameter to retune the model.

```
from sklearn.model_selection import GridSearchCV

param_grid = {
    'max_depth': [7, 8],
    'max_features': [11, 12, 13],
    'min_samples_leaf': [20, 25],
    'min_samples_split': [60, 75],
    'n_estimators': [101, 301]
}

rfcl = RandomForestRegressor()

grid_search = GridSearchCV(estimator = rfcl, param_grid = param_grid, cv = 3,n_jobs=-1)
```

We have trained the model on best parameters which are as follows:

```
{'max_depth': 8,
 'max_features': 13,
 'min_samples_leaf': 20,
 'min_samples_split': 60,
 'n_estimators': 301}
```

Adj R Square and RMSE values have been calculated and show the significance of the model. R square value for test shows 99.29 %, a model can predict the value of the target variable. We have other error also for the models.

Training Model Performance:

```
Mean Absolute Error (MAE): 728.8065923803891
Mean Squared Error (MSE): 965768.2292569826
Root Mean Squared Error (RMSE): 982.7350758251089
Mean Absolute Percentage Error (MAPE): 0.046749262116052244
R^2: 0.992877939362156
Adj R^2: 0.9928722369235973
```

Test Model Performance:

```
Mean Absolute Error (MAE): 734.2799283095094
Mean Squared Error (MSE): 937455.0075512985
Root Mean Squared Error (RMSE): 968.2226022724828
Mean Absolute Percentage Error (MAPE): 0.04668423709736856
R^2: 0.9929357720707594
Adj R^2: 0.9929301159372639
```

Hyperparameter tuning for RF model result is not improved much the performance metrics results. Based on all the build model, Random Forest is the best model for the problem statement. Performance matrices result shows RF model is better in performance compare to decision tree, bagging and linear regression. If we use hypertuning on RF model, we can achieve better result. For now, based on system constraint much better result is not achieved from the hyper tuned model.

We can use Random Forest model for prediction of the product_wgh_ton values based on the information of other values. We have seen storage issue reported l3m and temp reg mach are highly positive coefficient value, and location_type, transport_issue_l1y , wh_breakdown_l1m are highly negative coefficient value. These coefficients will play an essential role for the target variable. Based on our result, we have found '1684.32' value of product_wgh_ton is base value for the target variable. We have also found that if we dropped variables from the dataset which contains p value more than 0.05 than statsmodel performance can be improved.

6. Final interpretation / recommendation

In a Business prospective, , storage_issue_reported_l3m is highly correlated with product wg ton. So, We might need to increase the storage capacity of the warehouse to justify the demand in the respective zone.

Many warehouse has small capacity which could be utilize in a alternative to support the capacity of large or medium warehouse in demand of time.

In a North zone, regional zone 6 have the highest product wg ton requirement, so could increase the some more warehouse qty. We can circulate more offer in this area.

We have to be more attentive for zone 4 and zone 5 to bring it for more demand and supply balances

Based on VIF Analysis, retail_shop_num plays a most significant role in the variables importance. In a next view, Competitor_in_mkt, distributor_num , WH_regional_zone and Zone are the also a important factor variable.

Based on all the build model, Random Forest is the best model for the problem statement. Performance matrices result shows RF model is better in performance compare to decision tree, bagging and linear regression.

Hyper tuning of RF does not provide better result for our model.

Appendix:

Tableau Chart Plots:

https://public.tableau.com/app/profile/vinit.sharma2261/viz/Bivariateanalysisprojectnote1/WHCapacitywithproductwgton?publish=yes

https://public.tableau.com/app/profile/vinit.sharma2261/viz/Clusters_Analysis_project_notes_1/Clusterswhregionalwithproductandwarehousecapacity?publish=yes