Business Report

SMDM Project Business Report DSBA

*Capstone project – Supply Chain Management*

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 03-07-2022***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1** | **Introduction** | 5 |
| 2 | **EDA and business implication** | 6 |
| 3 | **Data cleaning and pre-processing** | 15 |
| 4 | **Model building** | 17 |
| 5 | **Model validation** | 21 |
| **6** | **Final interpretation / recommendation** | 22 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Sample Dataset. | 6 |
| 1.2 | Shape of the Dataset. | 7 |
| 1.3 | Description of the data. | 7 |
| 1.4 | Variable info of the dataset. | 7 |
| 1.5 | Univariate Plot | 9 |
| 1.6 | Count plot for approved\_wh\_govt\_certificate | 9 |
| 1.7 | Count plot for Location\_type | 10 |
| 1.8 | Count plot for WH\_capacity\_size | 10 |
| 1.9 | Count plot for Zone | 10 |
| 1.10 | Count plot for Warehouse in regional zone | 11 |
| 1.11 | Count plot for Warehouse owner type | 11 |
| 1.12 | Product Shipped across zone | 11 |
| 1.13 | Transport issue of shipment across zon | 12 |
| 1.14 | Storage issue of shipment across zone | 12 |
| 1.15 | Storage issue across product weight ton | 12 |
| 1.16 | Correlation data | 13 |
| 1.17 | Heatmap of Correlation plot | 13 |
| 1.18 | Pairplot | 14 |
| 1.19 | Missing values present in the data | 15 |
| 1.20 | Shape of the dataset after variable transformation. | 15 |
| 1.21 | sample dataset after variable transformation. | 15 |
| 1.22 | Info of the dataset after variable transformation and missing value treatment | 16 |
| 1.23 | Random Forest Regressor model initializing | 17 |
| 1.24 | Random Forest Regressor model best parameters | 17 |
| 1.25 | Random Forest Regressor model best Estimators | 17 |
| 1.26 | Random Forest Regressor model R-Square value | 17 |
| 1.27 | Random Forest Regressor model MSE and RMSE value | 17 |
| 1.28 | Random Forest Regressor model MAE value | 17 |
| 1.29 | Random Forest Regressor model MAPE value | 17 |
| 1.30 | ADA Boosting Regressor model Intializing | 18 |
| 1.31 | ADA Boosting Regressor model best parameters | 18 |
| 1.32 | ADA Boosting Regressor model best Estimators | 18 |
| 1.33 | ADA Boosting Regressor model R-Square value | 18 |
| 1.34 | ADA Regressor model MSE and RMSE value | 18 |
| 1.35 | ADA Boosting Regressor model MAE value | 18 |
| 1.36 | ADA Boosting Regressor model MAPE value | 18 |
| 1.37 | Bagging Regressor model Intializing | 19 |
| 1.38 | Bagging Regressor model best parameters | 19 |
| 1.39 | Bagging Regressor model best Estimators | 19 |
| 1.40 | Bagging Regressor model R-Square value | 19 |
| 1.41 | Bagging Regressor model MSE and RMSE value | 19 |
| 1.42 | Bagging Regressor model MAE value | 19 |
| 1.43 | Bagging Regressor model MAPE value | 19 |
| 1.44 | Ridge Regressor model Intializing | 19 |
| 1.45 | Ridge Regressor model best parameters | 19 |
| 1.46 | Ridge Regressor model best Estimators | 19 |
| 1.47 | Ridge Regressor model R-Square value | 20 |
| 1.48 | Ridge Regressor model MSE and RMSE value | 20 |
| 1.49 | Ridge Regressor model MAE value | 20 |
| 1.50 | Ridge Regressor model MAPE value | 20 |
| 1.51 | ANN Regressor model Intializing | 20 |
| 1.52 | ANN Regressor model best parameters | 20 |
| 1.53 | ANN Regressor model best Estimators | 20 |
| 1.54 | ANN Regressor model R-Square value | 20 |
| 1.55 | ANN Regressor model MSE and RMSE value | 21 |
| 1.56 | ANN Regressor model MAE value | 21 |
| 1.57 | ANN Regressor model MAPE value | 21 |
| 1.58 | Sorting best model based on Least MSE | 21 |
| 1.59 | Feature Selection based on the best model | 22 |

# 1. Introduction

*Business Problem:*

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.

*Brief introduction about the problem statement and the need of solving it:*

Brief introduction about the problem statement

The objective of this report is to find that, how the machine learning model supports the supply chain to overcome the demand and supply mismatch in every zone and warehouse. A FMCG company has entered into the instant noodles business two years back. The data is gathered based on the FMCG Company’s demand and supply mismatch of the product instant noodles. The higher management has noticed that there is a mismatch in the demand and supply of instant noodles.

The demand and supply mismatch can be overcome by following these:

* First of all, finding the demand and supply mismatch.
* Secondly, find the optimum weight of the product been shipped to each warehouse at different zone and regions of the country.

Need of solving it:

1. Company will lose heavily on logistic movement of goods / products
2. In order to sale the product, goods has to be moved where there is high supply or high demand zone.
3. Can minimize the inventory based on ROP (Reorder product) and ROQ (Reorder Quantity)
4. Profit of the company can be increased.
5. Stock [maintenance](https://www.google.com/search?sxsrf=ALiCzsYkXNNF5zwf92CCRITQ9TNvZvWfjg:1656780541086&q=maintenance&spell=1&sa=X&ved=2ahUKEwjK6vy81Nr4AhXu-jgGHTlmDOsQkeECKAB6BAgCEDM) in the inventory can be done.
6. Product quality can be improved.

2. EDA and Business Implication

Non visual Understanding of data:

The Dataset consist of 25000 rows of data with 22 independent variable and 1 target variable.

7 – Object type variable, 2 – float variable and 14 – integer type variable.

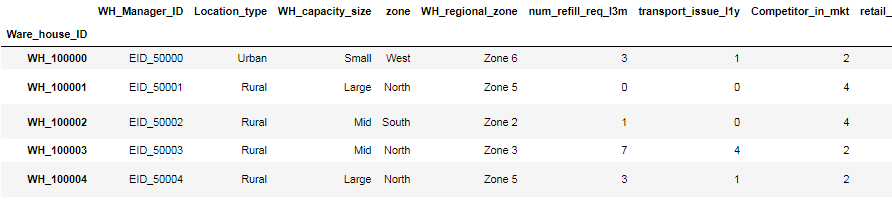


Fig 1.1 Sample Dataset.

Description of each and every variable in the dataset.

|  |  |
| --- | --- |
| **Variable** | **Business Definition** |
| 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 |
| product\_wg\_ton | Product has been shipped in last 3 months. Weight is in tons |



Fig 1.2 Shape of the Dataset.

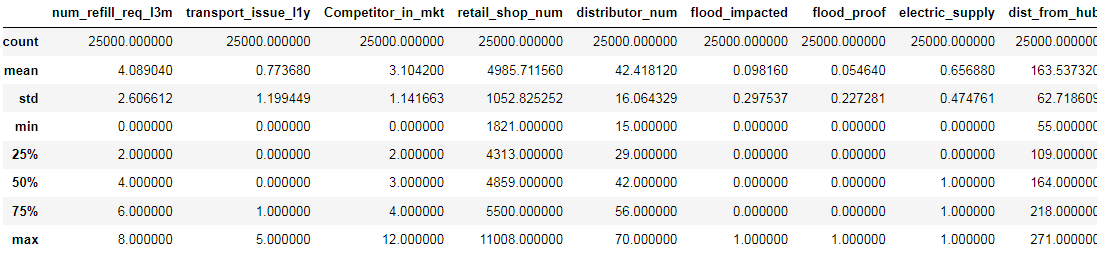


Fig 1.3 Description of the data.

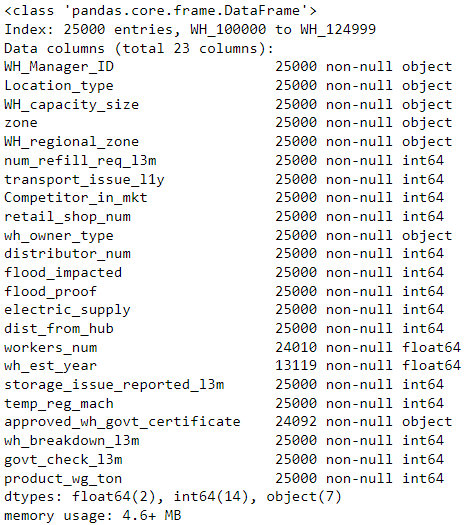


Fig 1.4 Variable info of the dataset.

Visual Understanding of data:

**Univariate Analysis:**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Fig 1.5 Univariate plot

**Product\_wg\_ton** – Target column is right skewed and there is no outliers present in the target column.

**Competitor\_in\_mkt -** Thisindependent variable has outlier in the dataset

**transport\_issue\_l1y -** Thisindependent variable has outlier in the dataset and right skewed values are present in the dataset.

**retail\_shop\_num -**  This independent variable has more outliers and values are slightly right skewed.

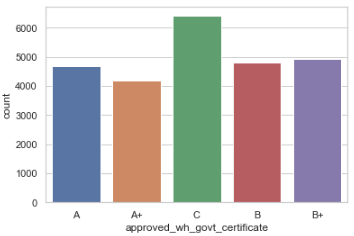


Fig 1.6 Count plot for approved\_wh\_govt\_certificate

From this count plot C certificate has the highest number of warehouse with government certificate A+ has the least number of warehouse government certificate.

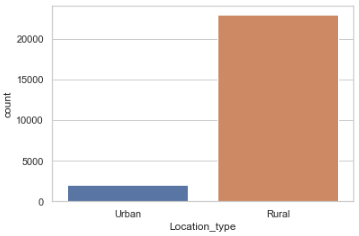


Fig 1.7 Count plot for Location\_type

Most number of ware houses is located in the rural area.

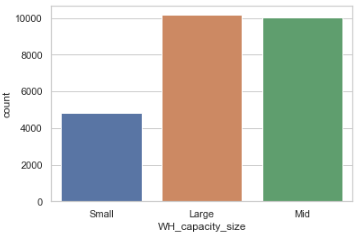


Fig 1.8 Count plot for WH\_capacity\_size

The Large capacity warehouse are having the ware house capacity higher than the other capacity ware houses.

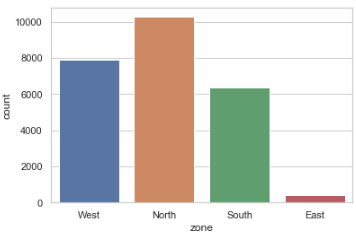


Fig 1.9 Count plot for Zone

North zone has the highest number of warehouse are built and East zone has the least number of ware houses.

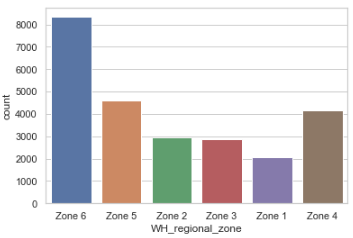


Fig 1.10 Count plot for Warehouse in regional zone



Fig 1.11 Count plot for Warehouse owner type

Most number of warehouse are owned by companies

**Bivariate analysis**

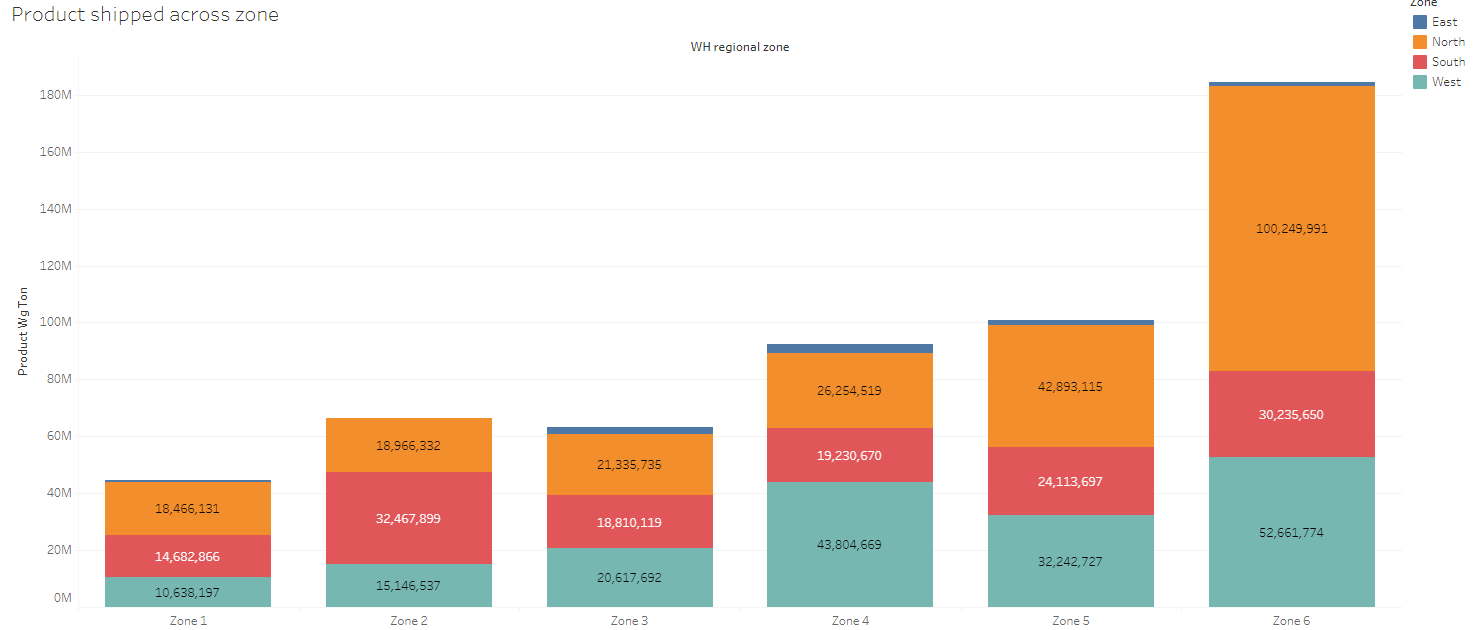


Fig 1.12 Product Shipped across zone

From this we can infer that product shipped in the east zone is very less. In North zone , Product shipping is higher every zone. We can infer that the supply is higher in the north zone and supply is lower in east zone.

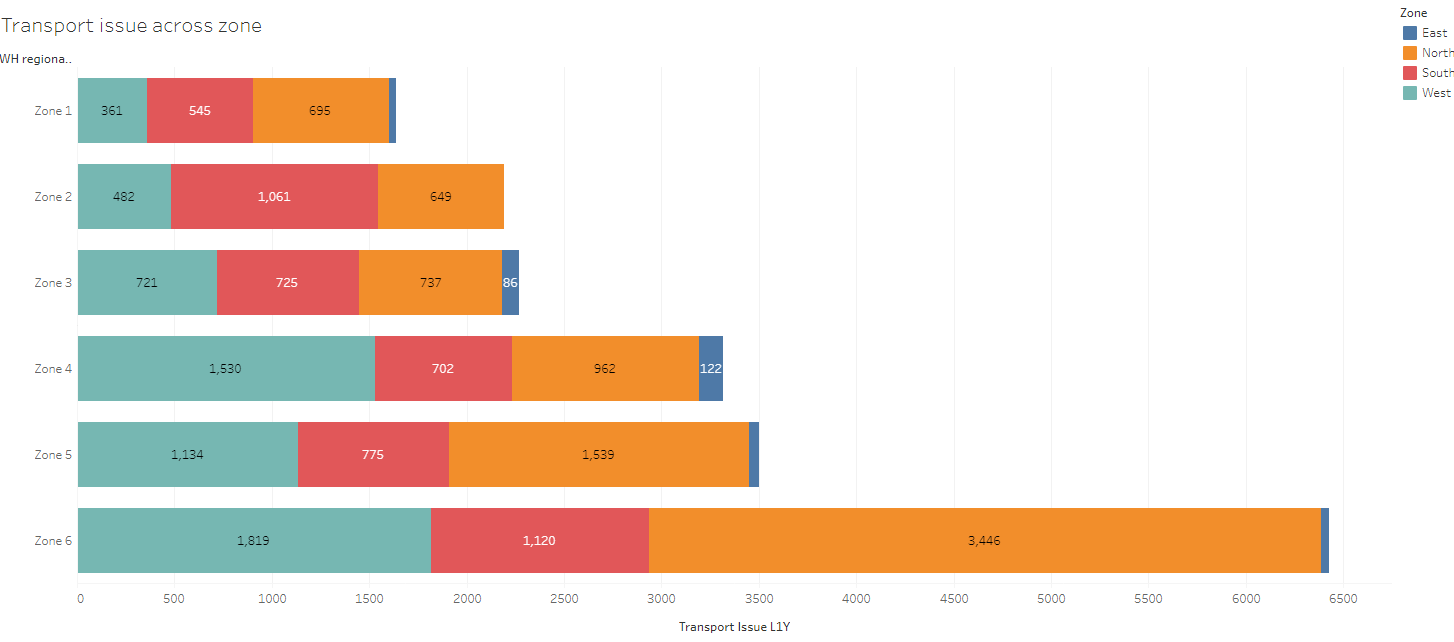


Fig 1.13 Transport issue of shipment across zone

Zone 6 has more number of transport issues when compared with other zone. In zone 6, maximum number of transport issue occurred in north zone.

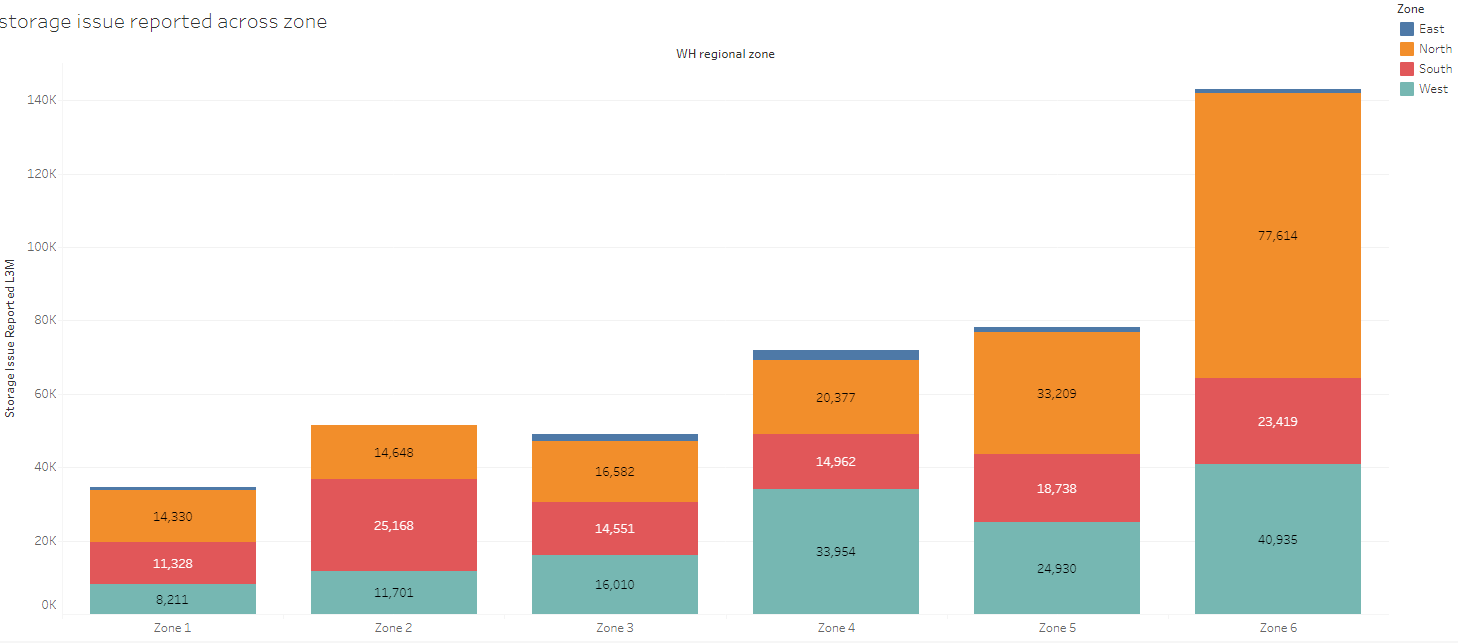


Fig 1.14 Storage issue of shipment across zone

The Zone -6 has the high supply products and high storage issues . The Storage area (Warehouse) has the limitation of stocking up the entire shipped product. Each Warehouse in every zone has the issue of storing issue in the warehouse.

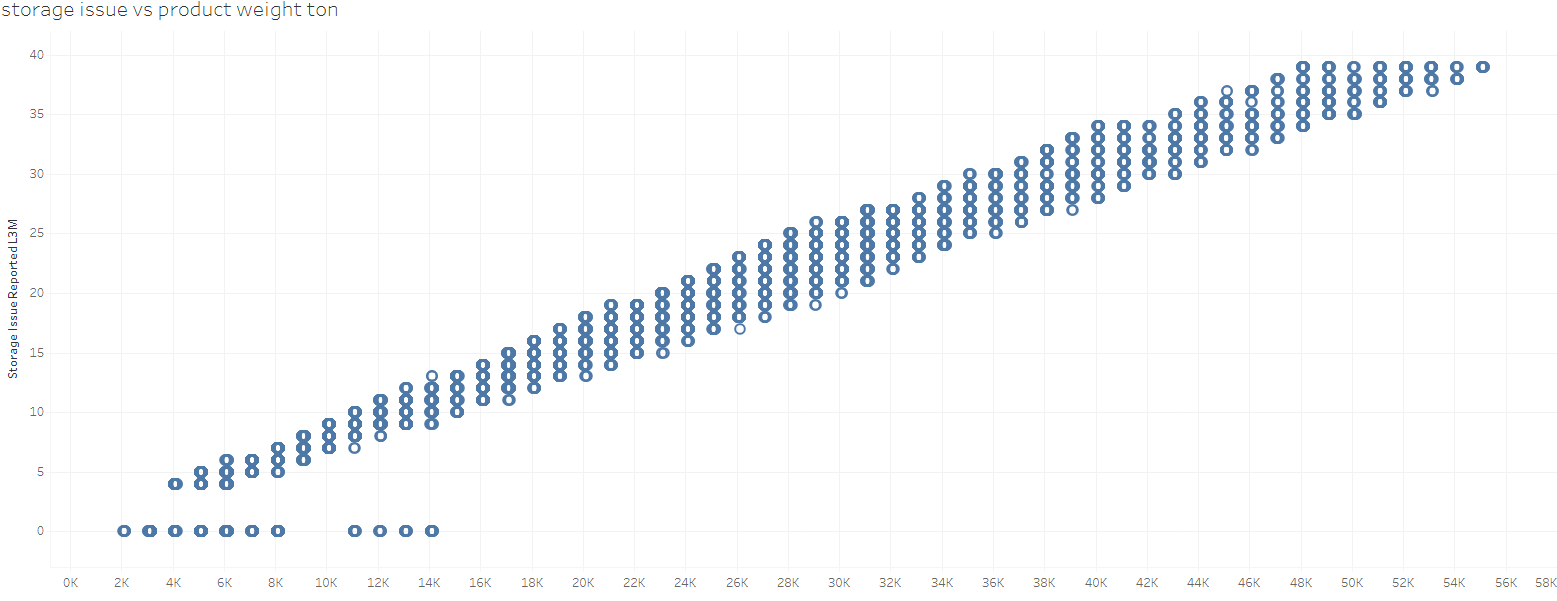


Fig 1.15 Storage issue across product weight ton

**Multivariate Analysis:**



Fig 1.16 Correlation data

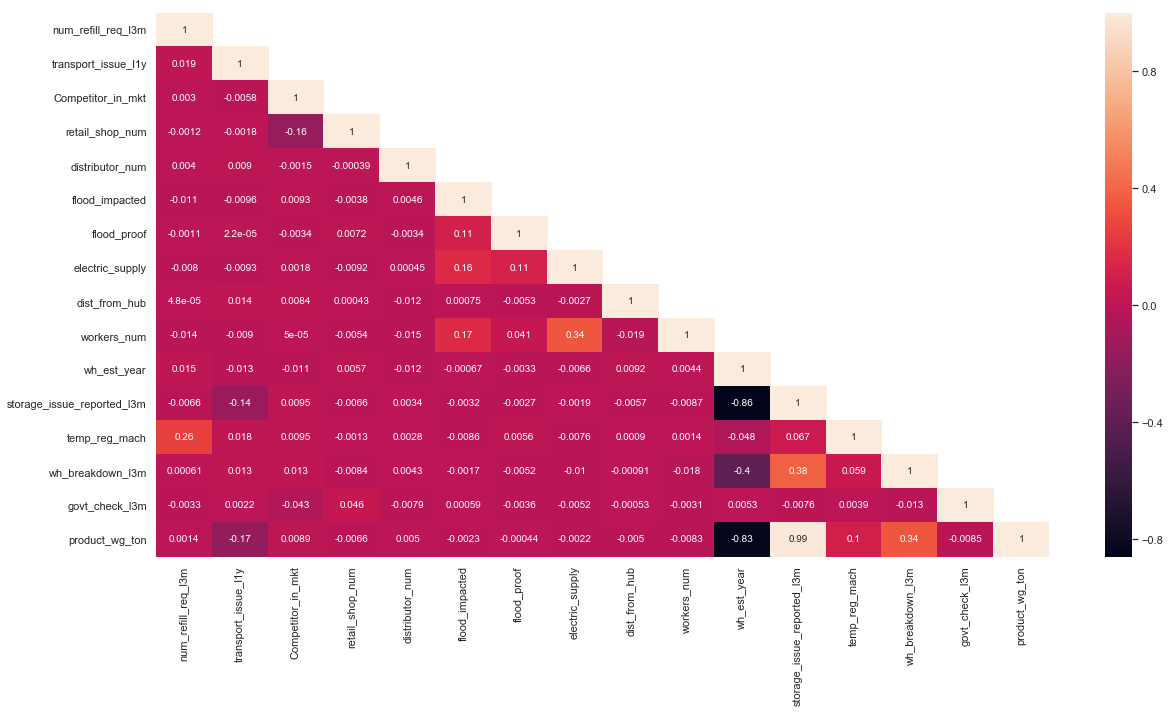
****

Fig 1.17 Heatmap of Correlation plot

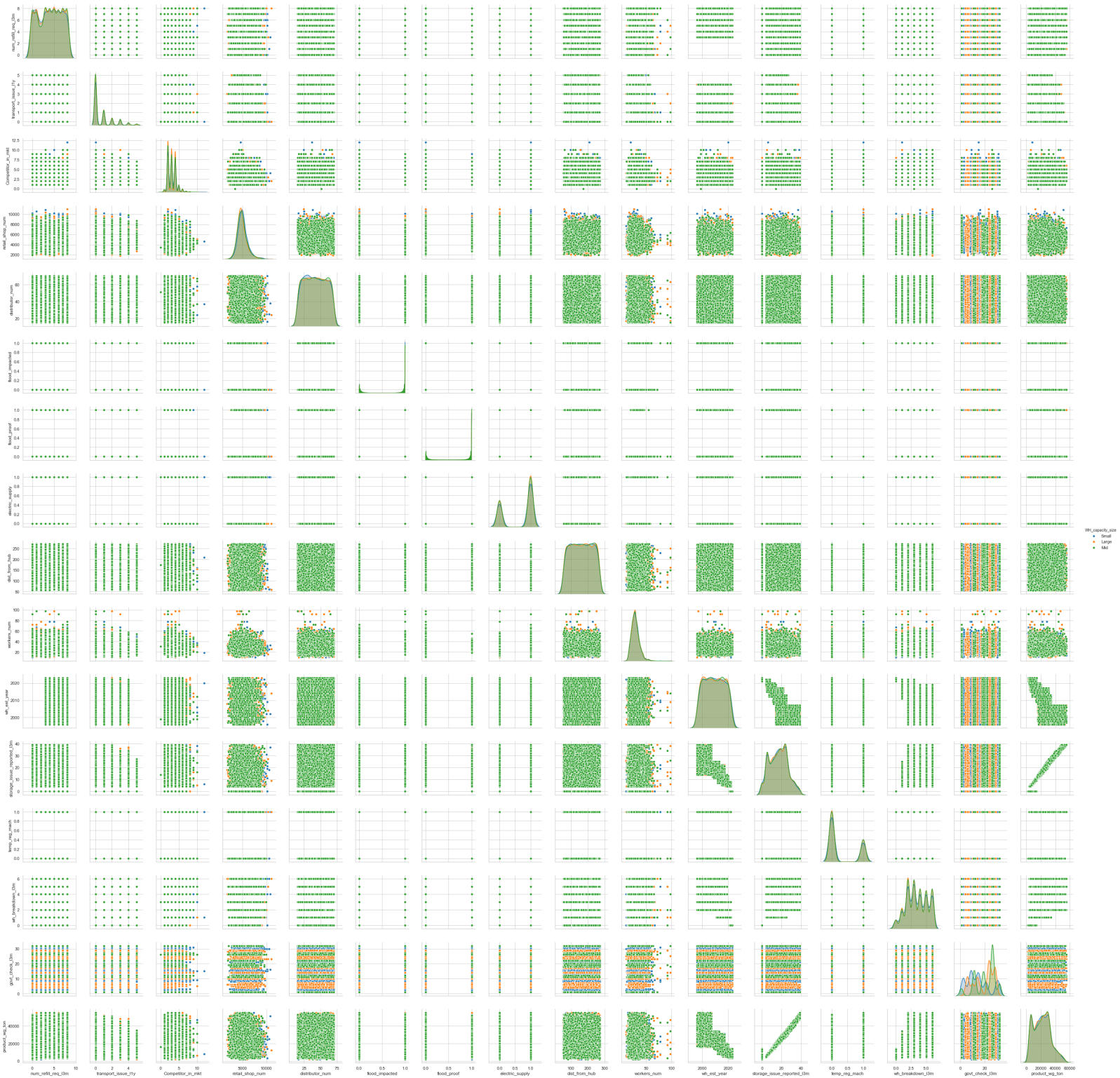
****

Fig 1.18 Pairplot

**How your analysis is impacting the business?**

From the above analysis, we can infer that, there is high supply in zone 6 when compared to other zone and Transport issues and storage issues are there in the every zone.

There is no correlation between the data.

# 3. Data Cleaning and Pre-processing

# *a.)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)*

|  |  |
| --- | --- |
|  |  |

Fig 1.19 Missing values present in the data

***Outlier Treatment:***

From the boxplot we can infer that, outlier treatment is not required for this dataset and discrete values are present in the data.

***Need for variable transformation (if any):***

Yes, Variable transformation is needed, because object type(text type) data are present in the data



Fig 1.20 Shape of the dataset after variable transformation.

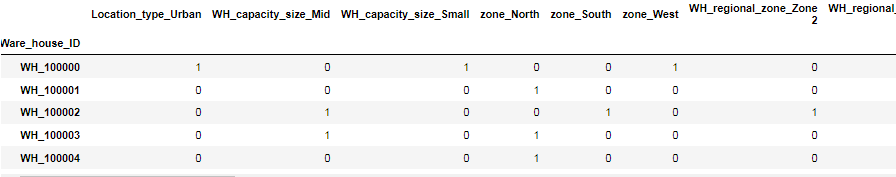


Fig 1.21 sample dataset after variable transformation.

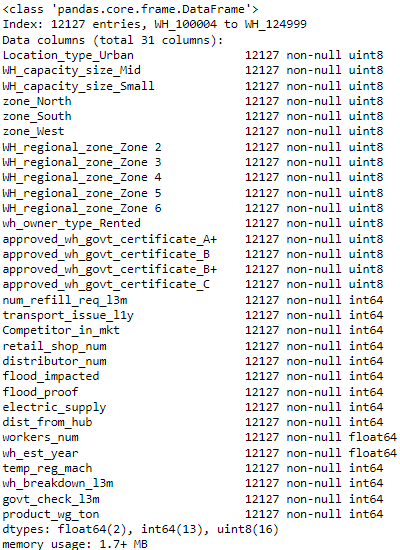


Fig 1.22 info of the dataset after variable transformation and missing value treatment

**Variables removed or added and why (if any):**

* As 'WH\_Manager\_ID' and 'Ware\_house\_ID' are unique values, we are dropping 'WH\_Manager\_ID' and setting 'Ware\_house\_ID' as an index value.
* As 'wh\_est\_year' having 48% of null values, so we are dropping 'wh\_est\_year' independent variable from the dataframe.
* 'storage\_issue\_reported\_l3m' independent variable is highly correlated with the target column. So 'storage\_issue\_reported\_l3m' can also be dropped from the dataframe.

**4**. **Model building**

*Clear on why was a particular model(s) chosen. Effort to improve model performance.*

***Clear on why was a particular model(s) chosen:***

***Random Forest Regressor model:***

* It reduces over fitting in decision trees and helps to improve the accuracy
* It is flexible to both classification and regression problems
* It works well with both categorical and continuous value.

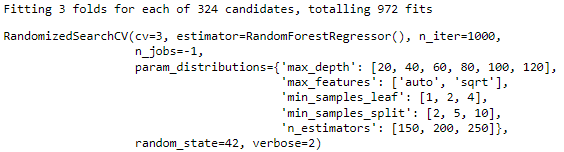


Fig 1.23 Random Forest Regressor model initializing

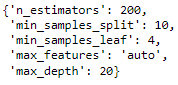


Fig 1.24 Random Forest Regressor model best parameters



Fig 1.25 Random Forest Regressor model best Estimators



Fig 1.26 Random Forest Regressor model R-Square value

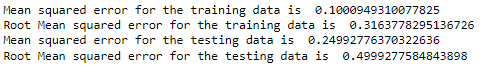


Fig 1.27 Random Forest Regressor model MSE and RMSE value



Fig 1.28 Random Forest Regressor model MAE value



Fig 1.29 Random Forest Regressor model MAPE value

### *Ada Boosting Regressor model:*

* Adaboost is less prone to overfitting as the input parameters are not jointly optimized.
* The accuracy of weak classifiers can be improved by using Adaboost.

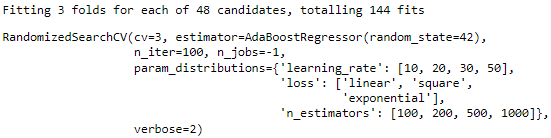


Fig 1.30 ADA Boosting Regressor model Intializing



Fig 1.31 ADA Boosting Regressor model best parameters



Fig 1.32 ADA Boosting Regressor model best Estimators



Fig 1.33 ADA Boosting Regressor model R-Square value

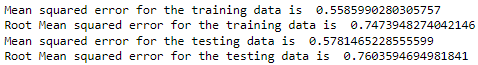


Fig 1.34 ADA Regressor model MSE and RMSE value



Fig 1.35 ADA Boosting Regressor model MAE value



Fig 1.36 ADA Boosting Regressor model MAPE value

### *Bagging Regressor model:*

* **Ease of implementation**
* **Reduction of variance**
* significantly raises the stability of models in improving accuracy

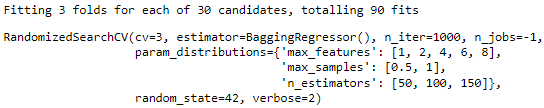


Fig 1.37 Bagging Regressor model Intializing



Fig 1.38 Bagging Regressor model best parameters



Fig 1.39 Bagging Regressor model best Estimators



Fig 1.40 Bagging Regressor model R-Square value

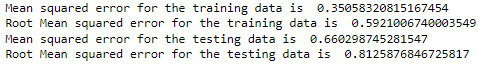


Fig 1.41 Bagging Regressor model MSE and RMSE value



Fig 1.42 Bagging Regressor model MAE value



Fig 1. 43 Bagging Regressor model MAPE value

### Ridge Regression model:

* Avoids overfitting model
* Performs well in large multivariate data.

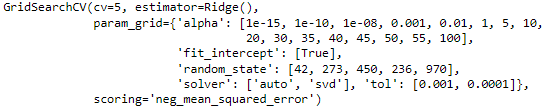


Fig 1.44 Ridge Regressor model Intializing

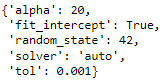


Fig 1.45 Ridge Regressor model best parameters



Fig 1.46 Ridge Regressor model best Estimators



Fig 1.47 Ridge Regressor model R-Square value

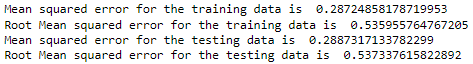


Fig 1.48 Ridge Regressor model MSE and RMSE value



Fig 1.49 Ridge Regressor model MAE value



Fig 1.50 Ridge Regressor model MAPE value

### ANN Regression model:

* Storing information on the entire network
* Having a distributed memory
* Parallel processing capability

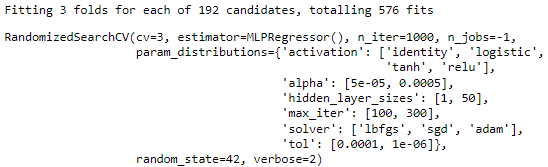


Fig 1.51 ANN Regressor model Intializing

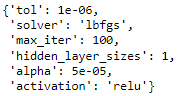


Fig 1.52 ANN Regressor model best parameters



Fig 1.53 ANN Regressor model best Estimators



Fig 1.54 ANN Regressor model R-Square value

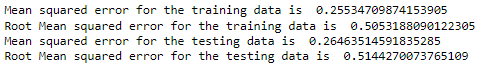


Fig 1.55 ANN Regressor model MSE and RMSE value



Fig 1.56 ANN Regressor model MAE value



Fig 1.57 ANN Regressor model MAPE value

***Effort to improve model performance.***

Hyper Parameter tuning has been performed for increasing the performance of the model.

**5. Model validation**

*How was the model validated? Just accuracy, or anything else too?*

The model was validated through calculating the multiple parameters for each model.

Parameters are

* R Squared value (R2 value)
* Mean Squared Error value (MSE)
* Root Mean Squared Error value (MSE)
* Mean Absolute Error value (MAE)
* Mean Absolute Percentage Error value (MAPE)

A good or best model needs to have the value in the range.

* RMSE - 0.2-0.5
* Mean absolute error (MAE) -10
* R2 value - near to 1
* Mean squared error (MSE) - near to 0

Best model is identified from the calculated value of each parameter for each model are:

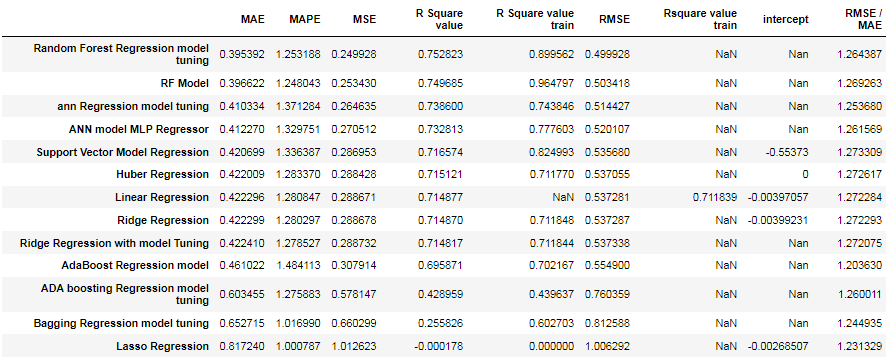


Fig 1.58 Sorting best model based on Least MSE

**6. Final interpretation / recommendation**

*Detailed recommendations for the management/client based on the analysis done.*

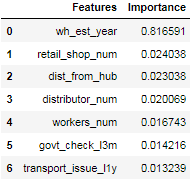


Fig 1.59 Feature Selection based on the best model

These are top 7 features that are important for this data.

**Wh\_est\_year** – During early periods the supply was higher and the year progress the shipping of product weight is reduced. May be population has been increased over the years and production/manufacturing of the product has not been increased.

**retail\_shop\_num** – The Availability of the products in the shops, needs to be increased. Because demand is higher and supply is less.

**dist\_from\_hub** – The warehouse may be located far from the manufacturing plant. This may take lot off time for transportation of goods and cost of transportation may increase.

***Insights and Recommendations:***

* Inventory can be managed based on the optimization of warehouse details.
* Transportation cost can be minimized.
* Supply of goods can be time delay.
* Better collaboration with suppliers.
* Better quality control.
* Shipping optimisation.
* Reduced inventory and overhead costs.
* Improved risk mitigation.
* Stronger cash flow.

***Recommendations:***

* Need to increase warehouse in the urban areas.
* East zone has the least warehouse and the demand and supply is low. Need to promote more about the product and offers in the East zone.
* Need to give more offers to the Dealers for buying products as bulk orders.
* Need to employee more workers in the warehouse for monitoring the inventory. Increasing of workers in warehouse can help the workers to monitor the product.
* Need to give more offers to the customer based on the zone to sell more products as the competitor are there in the market.
* Minimum order and Maximum orders needs to be implemented for the Dealers or wholeseller for the Quality of the product.