Business Report

SMDM Project Business Report DSBA

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 15-05-2022***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1)** | **Introduction of the business problem** | 4 |
|  | a) Defining problem statement | 4 |
|  | b) Need of the study/project | 4 |
|  | c) Understanding business/social opportunity | 5 |
|  |  | 5 |
| **2)** | **Data Report** | 6 |
|  | a) Understanding how data was collected in terms of time, frequency and methodology | 6 |
|  | b) Visual inspection of data (rows, columns, descriptive details) | 10 |
|  | c) Understanding of attributes (variable info, renaming if required) | 11 |
|  |  | 15 |
| **3)** | **Exploratory data analysis** | 19 |
|  | a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) | 23 |
|  | b) Bivariate analysis (relationship between different variables , correlations) | 24 |
|  | c) Removal of unwanted variables (if applicable) | 28 |
|  | d) Missing Value treatment (if applicable) | 30 |
|  | d) Outlier treatment (if required) | 33 |
|  | e) Variable transformation (if applicable) | 35 |
|  | f) Addition of new variables (if required) | 36 |
|  | g) Addition of new variables (if required) |  |
|  |  |  |
| **4)** | **Business insights from EDA** |  |
|  | a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business |  |
|  | b) Any business insights using clustering  (if applicable) |  |
|  | c) Any other business insights |  |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Dataset Sample Before Changing Column Names | 5 |
| 1.2 | Dataset Sample After Changing Column Names | 5 |
| 1.3 | Sample Datatypes of the variable with null values | 6 |
| 1.4 | Shape before Outliers Treatment | 7 |
| 1.5 | Shape After Outliers Treatment | 7 |
| 1.6 | Before Treating Missing value | 8 |
| 1.7 | After Treating Missing value | 8 |
| 1.8 | Default count. | 9 |
| 1.9 | Default count in percentage. | 9 |
| 1.10 | Sample data after Transformation. | 9 |
| 1.11 | Univariate Analysis | 9 |
| 1.12 | Scatterplot for Bivariate Analysis | 10 |
| 1.13 | Sample Multivariate analysis for correlation | 11 |
| 1.14 | Multivariate analysis of plotting correlation in heatmap | 12 |
| 1.15 | Multivariate analysis of plotting correlation in heatmap after dropping insignificant variable. | 13 |
| 1.16 | Taking Highly correlated variables. | 13 |
| 1.17 | X-train shape. | 13 |
| 1.18 | X-test shape. | 13 |
| 1.19 | Y-train shape. | 13 |
| 1.20 | Y-test shap | 13 |
| 1.21 | Selecting the feature with rank 1 | 14 |
| 1.22 | Model-1 summary report | 14 |
| 1.23 | Variance Inflation Factor.of Model-1 | 15 |
| 1.24 | Model-2 Summary Report | 15 |
| 1.25 | Model-3 Summary Report | 16 |
| 1.26 | Model-4 Summary Report | 17 |
| 1.27 | Model-5 Summary Report | 18 |
| 1.28 | Model-6 Summary Report | 19 |
| 1.29 | Model-7 Summary Report | 20 |
| 1.30 | Model-8 Summary Report | 21 |
| 1.31 | Model-9 Summary Report | 22 |
| 1.32 | Optimum threshold | 22 |
| 1.33 | Confusion matrix for train data | 22 |
| 1.34 | Confusion matrix for test data | 22 |
| 1.35 | Classification report for train data | 23 |
| 1.36 | Classification report for test data | 23 |
| 1.37 | Initializing Random Forest Classifier | 24 |
| 1.38 | Taking features with Rank 1 | 24 |
| 1.39 | value count of the target column. | 24 |
| 1.40 | Initializing Model-1 using RF model | 24 |
| 1.41 | Model-1 Summary. | 25 |
| 1.42 | Model-1 Variation Inflation Matrix (VIF) | 25 |
| 1.43 | Initializing Model-2 | 26 |
| 1.44 | Model-2 Summary. | 26 |
| 1.45 | Initializing Model-3 | 26 |
| 1.46 | Model-3 Summary | 27 |
| 1.47 | Initializing Model-4 | 27 |
| 1.48 | Model-4 Summary | 28 |

|  |  |  |
| --- | --- | --- |
| 1.49 | Boxplot for Default variable. | 28 |
| 1.50 | Optimum threshold value | 28 |
| 1.51 | Predicted train values | 29 |
| 1.52 | Predicted test values | 29 |
| 1.53 | Boxplot for test values | 29 |
| 1.54 | Optimum threshold values for test data | 29 |
| 1.55 | Confusion matrix train values | 30 |
| 1.56 | Confusion matrix test values | 30 |
| 1.57 | Classification report for train data | 30 |
| 1.58 | Classification report for test data | 30 |
| 1.59 | ROC for train data | 30 |
| 1.60 | AUC score for train data | 30 |
| 1.61 | ROC for test data | 30 |
| 1.62 | AUC score for test data | 30 |
| 1.63 | Initializing LDA model | 30 |
| 1.64 | Taking features with rank 1 for LDA model | 30 |
| 1.65 | Value count for default variable | 31 |
| 1.66 | Model 1 Initializing | 31 |
| 1.67 | Model -1 summary | 31 |
| 1.68 | Model -1 VIF | 32 |
| 1.69 | Model – 2 Intializing | 32 |
| 1.70 | Model – 2 Summary | 32 |
| 1.71 | Intializing Model -3 | 33 |
| 1.72 | Model -3 Summary | 33 |
| 1.73 | Default value for LDA train model | 33 |
| 1.74 | Optimum threshold value for LDA train model | 34 |
| 1.75 | Predicted value for LDA train model | 34 |
| 1.76 | Default value for LDA test model | 34 |
| 1.77 | Predicted value for LDA test model | 34 |
| 1.78 | Confusion matrix train values | 35 |
| 1.79 | Confusion matrix test values | 35 |
| 1.80 | Classification report for train data | 35 |
| 1.81 | Classification report for test data | 35 |
| 1.82 | ROC for train data | 35 |
| 1.83 | AUC score for train data. | 35 |
| 1.84 | ROC for test data | 35 |
| 1.85 | AUC score for test data. | 35 |
| 1.86 | Comparison dataframe for LR,RF and LDA values. | 35 |
| 1.87 | ROC curve for LR model | 36 |
| 1.88 | ROC curve for RF model | 36 |
| 1.89 | ROC curve for LDA model | 36 |

# Introduction of the business problem

*a) Defining 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. 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.

Drawback of demand and supply mismatch:

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.

*b) Need of the study/project:*

1. To find the mismatch in demand and supply so that management can optimize the supply quantity in each and every warehouse in entire country.
2. The optimum weight of the product shipped every time based on the demand on each zone or region.
3. Data analysis will help us to analysis the product quantity sale based on the zone wise demand.
4. Loss can be minimized by the management based on data analysis.

To meet demand and supply:

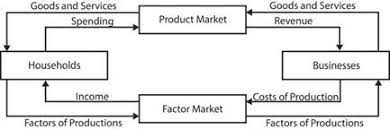
1. Promotion of products through advertisements in low demanding zone.
2. New marketing strategies to be implemented based on low and high demand of product zone. Marketting strategies are:

* Posters
* Pamphlets
* Offers
* Sign Boards
* Health benefits of the product

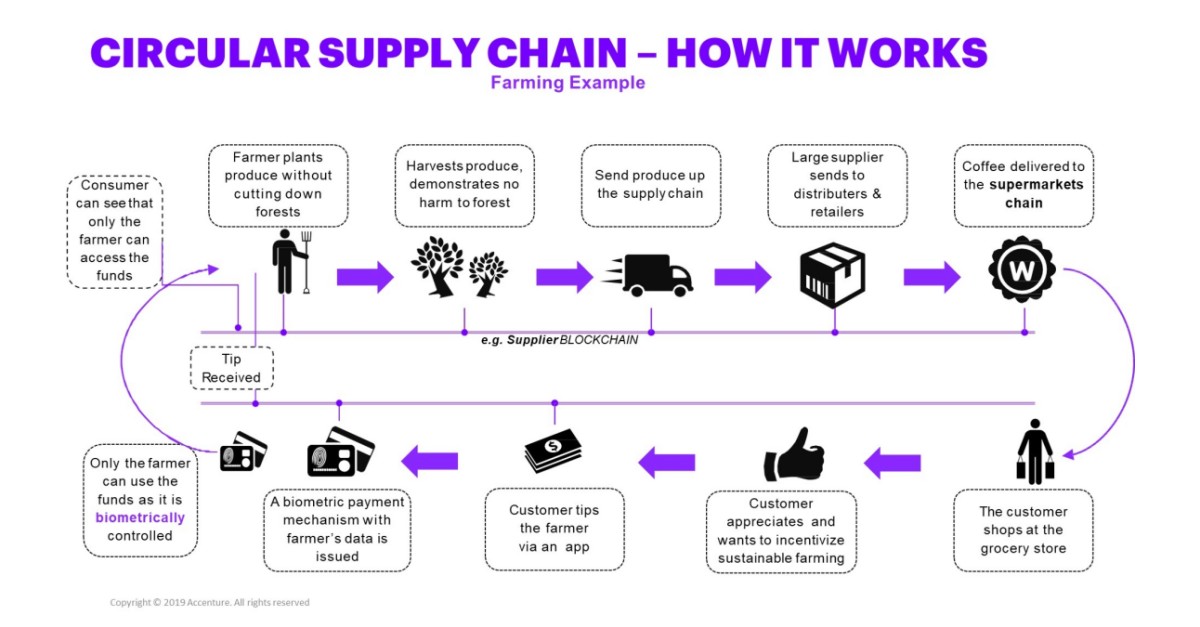
*c) Understanding business/ social opportunity:*

1. The demand patterns of the customers across different zones can help the company in understanding the customer behaviour and needs of the customer in every zone.
2. By this, we can forecast the demand of the products in every zone and needed quantity can be shipped to every dealer/ retailer based on the demand in every zone (Resources in demand & shipping of the product).
3. The company can monitor the performance of the goods/product in every zone and offers based on low and high selling zone.
4. The optimum maintenance and cash margin can be given based on the demand of the product in every zone.
5. Based on the demand of the product, the stocks can be refilled in the warehouses in every zone.
6. Rotation of Stocks regularly helps the company to reduce the loss, improves the efficiency and product demand can be improved.
7. This helps the business to optimize the performance.

### Without Supply chain and logistics:



**With Supply chain and logistics:**

****

**Data Report**

*a) Understanding how data was collected in terms of time, frequency and methodology.*

***Methodology:***

7 Data Collection Methods Used in Business Analytics are :

1. Surveys
2. Transactional Tracking
3. Interviews and Focus Groups
4. Observation
5. Online Tracking
6. Forms
7. Social Media Monitoring

Transactional Tracking and Survey type of method are used in the data collection of this dataset for the FMCG Company.

***Time:***

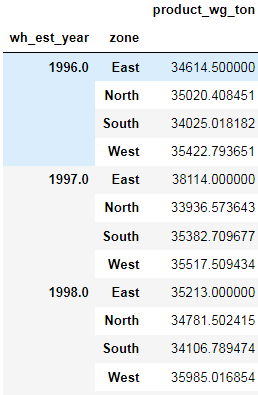


Fig 1.1 Sample data of product weight across years

In the year **1996**, the demand of the product is higher in the **west** **zone** when compared with the other zone.

***Frequency:***

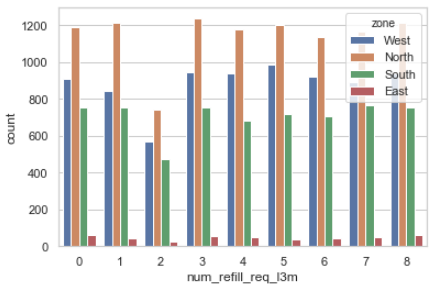


Fig 1.2 Number of times frequently product refilled across zone

North zone has the highest number of times the stock / product has been refilled. From this we can infer that, North zone has high demand of the product.

*b) Visual inspection of data (rows, columns, descriptive details):*

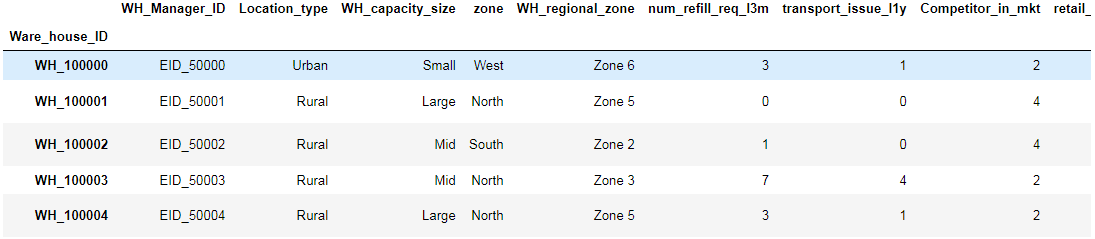


Fig 1.3 Sample dataset

Dataset consist of 25000 rows and 23 columns. 22 are independent variable and 1 target variable.

‘Product\_wg\_ton’ is the target column.

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 1.4 Shape of the dataset

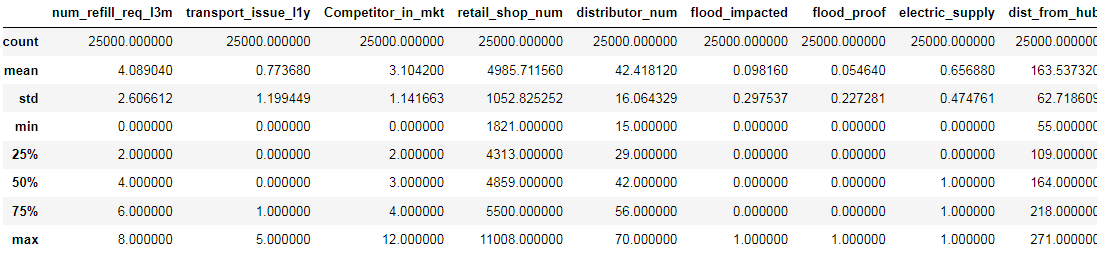


Fig 1.5 Description of the data.

*c) Understanding of attributes (variable info, renaming if required):*

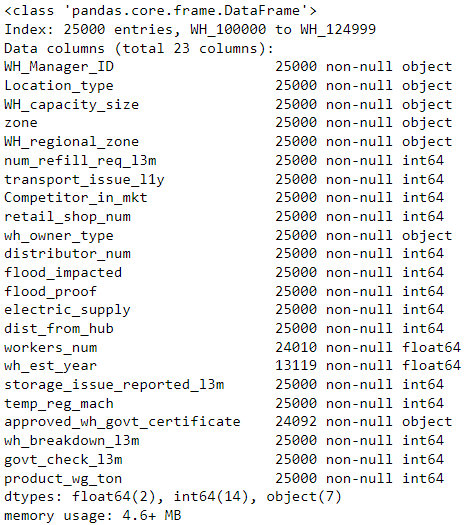


Fig 1.6 Variable info of the dataset.

From this we can infer that, 7 variable are object type, 14 variable are integer type and 2 are float type variable. 3 variable has some missing value in the dataset.

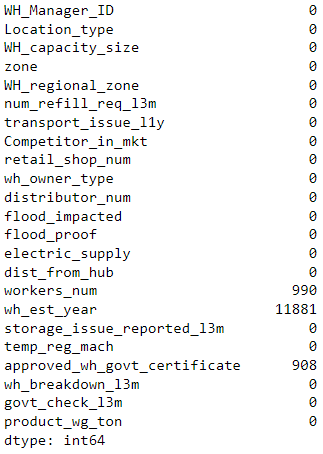


Fig 1.7 Missing value count in the dataset.

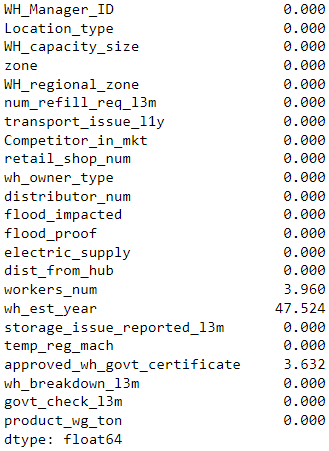


Fig 1.8 Missing value count in percentage.

Around **48 % (11881)** missing values data are present in the “**wh\_est\_year**” variable.

Renaming of the variable is not required for this dataset.

**Exploratory data analysis**

*a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones):*

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

Fig 1.9 Univaraite Analysis

**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.

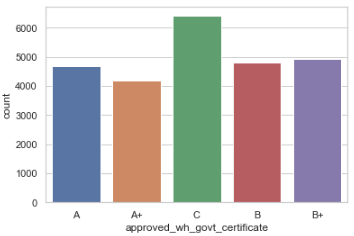


Fig 1.10 Count plot for approved\_wh\_govt\_certificate

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

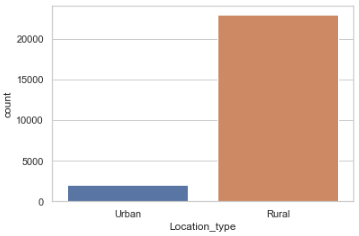


Fig 1.11 Count plot for Location\_type

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

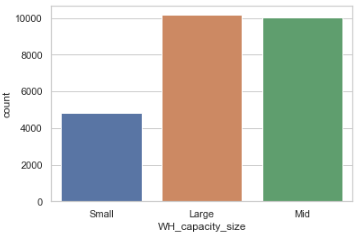


Fig 1.12 Count plot for WH\_capacity\_size

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

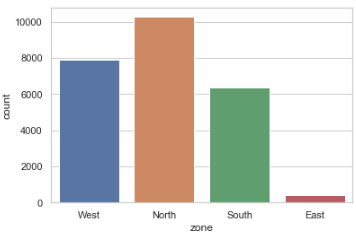


Fig 1.13 Count plot for Zone

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

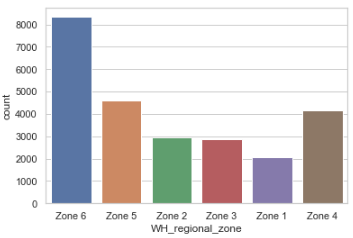


Fig 1.14 Count plot for Warehouse in regional zone



Fig 1.15 Count plot for Warehouse owner type

Most number of warehouse are owned by companies

*b) Bivariate analysis (relationship between different variables , correlations):*

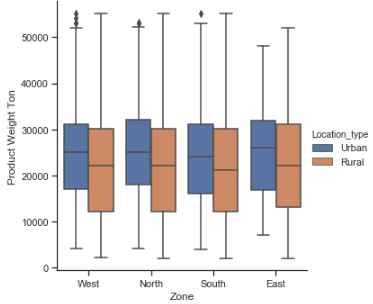


Fig 1.16 boxplot for Product weight ton vs. zone across location type.

The demand for the product in the urban areas is higher when compared with the rural area. Outliers are present in the urban area, the inference from the above boxplot.

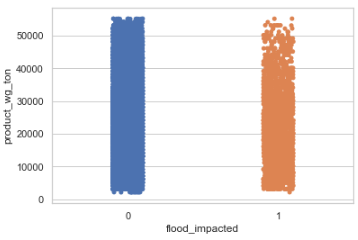


Fig 1.17 stripplot for Product weight ton vs. flood\_impacted.

The Flood indicator indicates that most of the products in the ware house are affected by the flood.

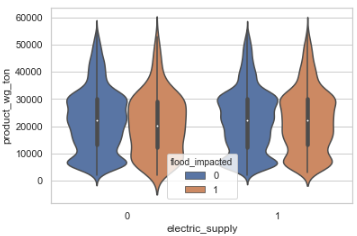
.

Fig 1.18 Violinplot for Product weight ton vs. electric\_supply with flood\_impacted.

The warehouse is having the backup power supply where the flood impacted area will have the loss of power or

when there is power shutdown in the area.

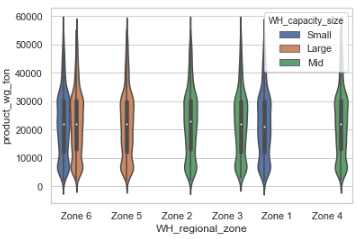


Fig 1.19 Violinplot for Product weight ton vs. wh\_regional\_zone with wh\_capacity\_size.

2 types of ware house capacity size (small and large) are there in zone 6 and other zone has either one type of ware house capacity (small, medium, large).

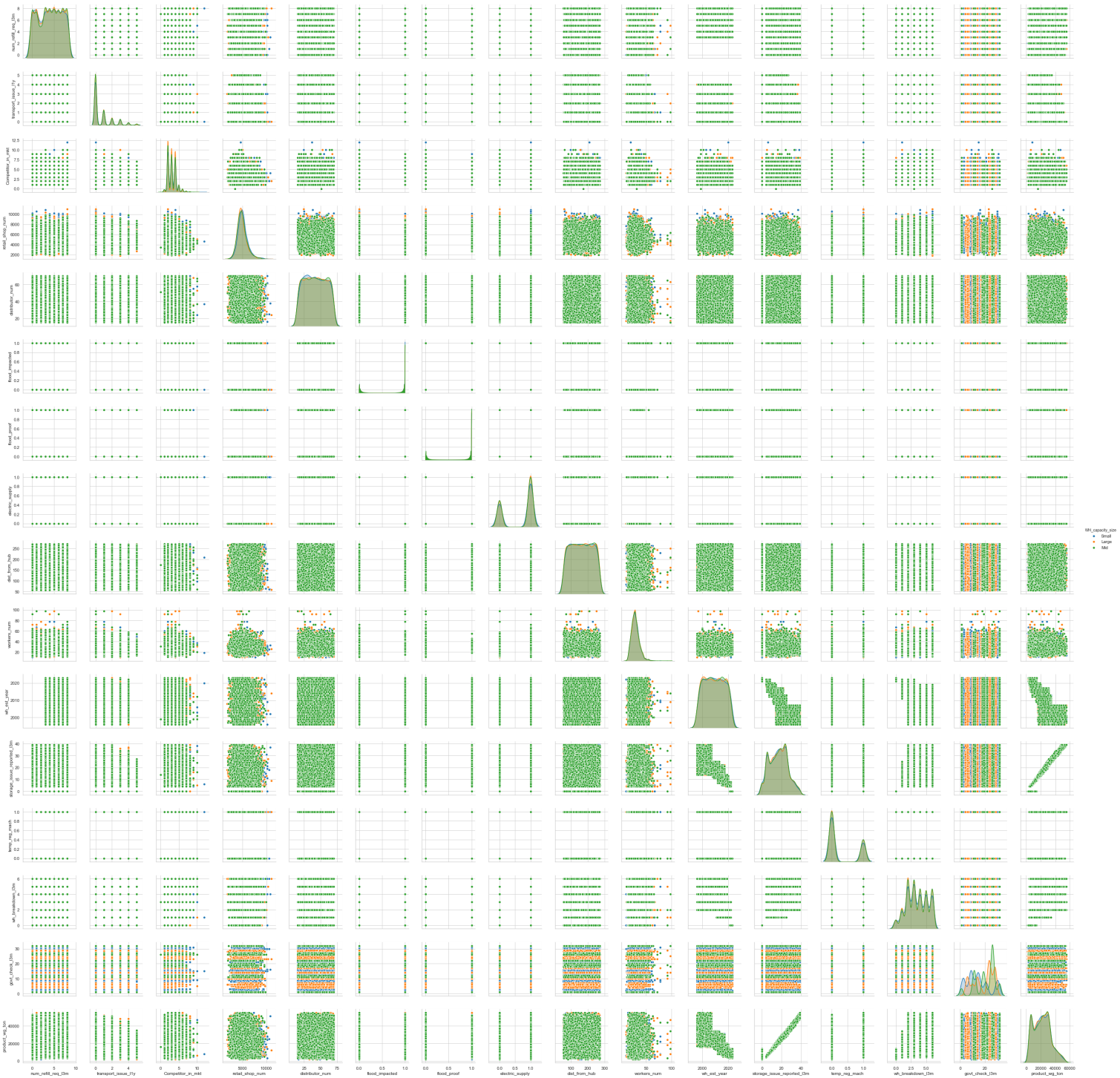


Fig 1.19 Pairplot for the bivariate analysis.

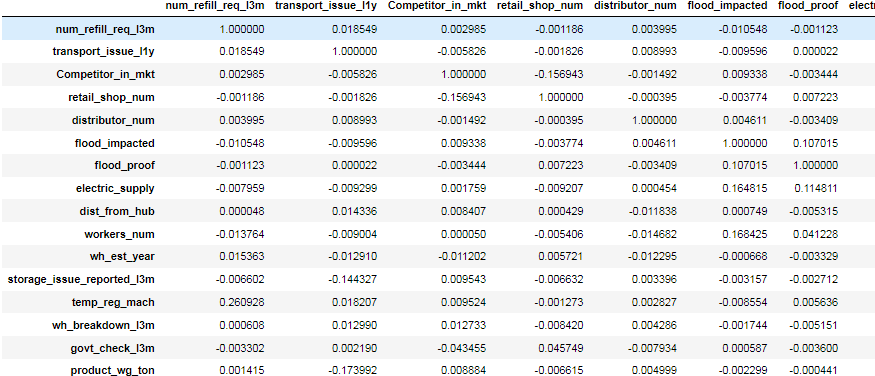


Fig 1.20 sample Correlation data for the bivariate analysis.

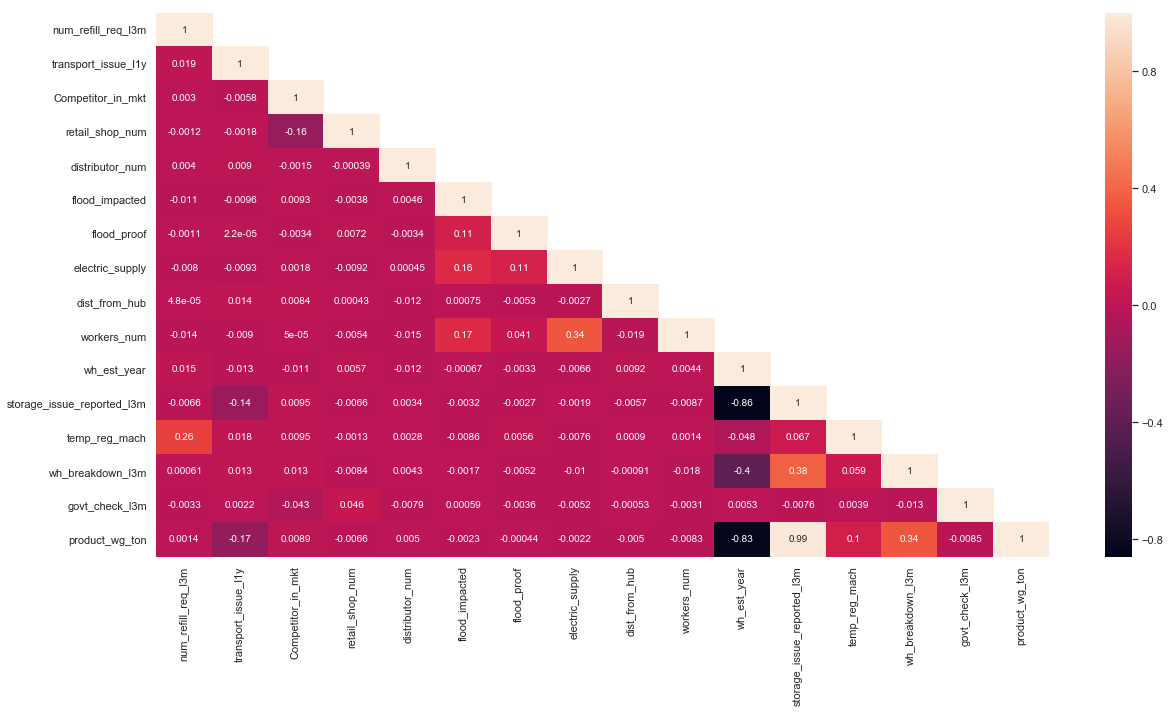


Fig 1.21 Plotting correlation in the heatmap for the bivariate analysis.

From the heat map, we can infer that, ‘**storage\_issue\_reported\_l3m**’ (storage issue reported last 3 months) is **highly correlated** with the target variable (‘**product\_wg\_ton’**).

Wh\_est\_year has **highly negatively correlated** with the target variable (‘**product\_wg\_ton’**) and **storage\_issue\_reported\_l3m’** variable

*c) Removal of unwanted variables (if applicable):*

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**' is 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.



Fig 1.22 shape of the dataframe after dropping unwanted variables

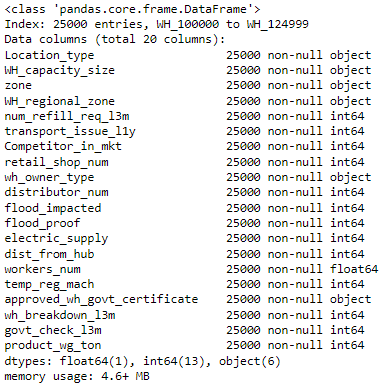


Fig 1.23 Info of the dataframe after dropping unwanted variables.

### *d) Missing Value treatment (if applicable):*

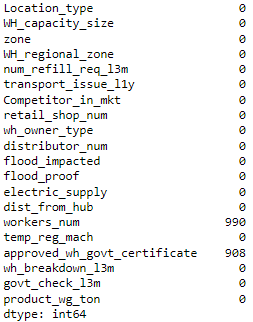


Fig 1.24 Missing values before treating.

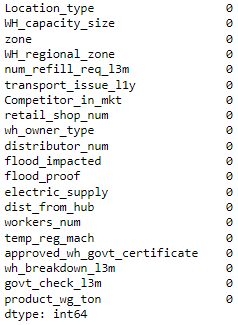


Fig 1.25 Missing values after treating

Missing values are treated using mean value of integer type and mode value for object type variable and there are no missing values present in the dataset.

### *e) Outlier treatment (if required):*

From the boxplot we can infer that, outlier treatment is not required for this dataset.

*f) Variable transformation (if applicable):*



Fig 1.26 Shape of the dataset after variable transformation.

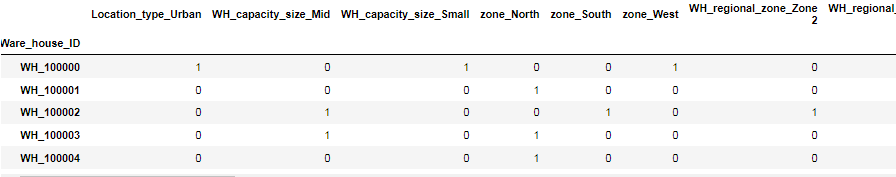


Fig 1.27 sample dataset after variable transformation.

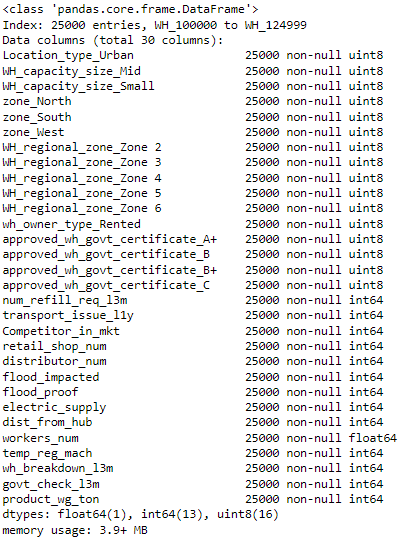


Fig 1.28 Info of the dataset after variable transformation.

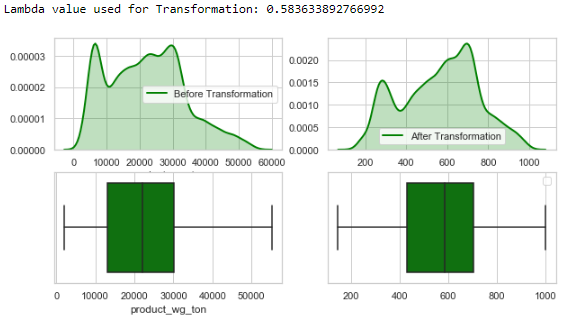


Fig 1.29 Transforming right skewed target variable.

The target variable is right skewed before transformation. After transforming the target variable the values are now changed to a slightly normal distributed value.

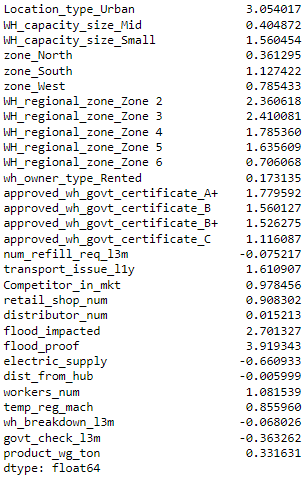


Fig 1.30 Finding skewness of the data frame.

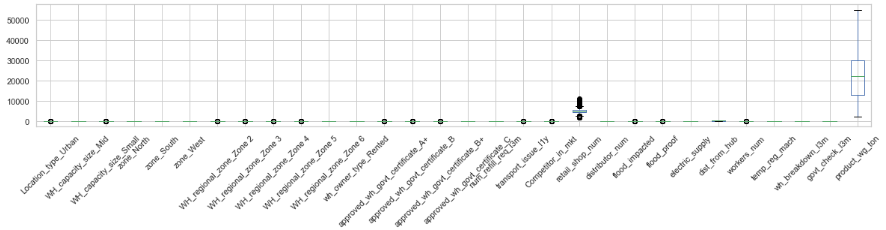
.

Fig 1.31 Boxplot before scaling

*g) Addition of new variables (if required):*

Addition of new variables is not required in this dataset.

**Business insights from EDA**

*a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business:*

If the data is unbalanced, we can use Smote-R and Smote-R Gaussian for treating the unbalanced data.

*b) Any business insights using clustering (if applicable)*

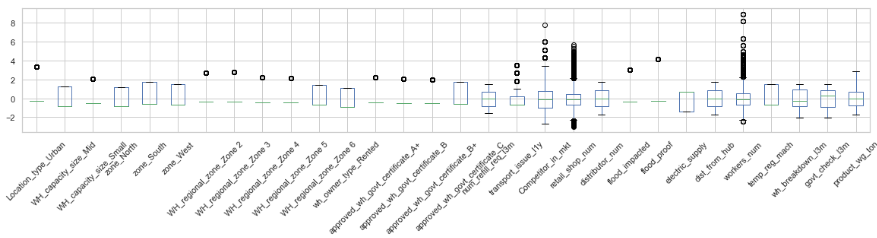


Fig 1.32 Boxplot after scaling

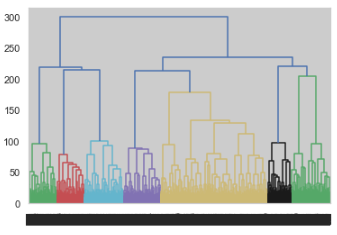


Fig 1.33 Dendogram of hierarchial clustering

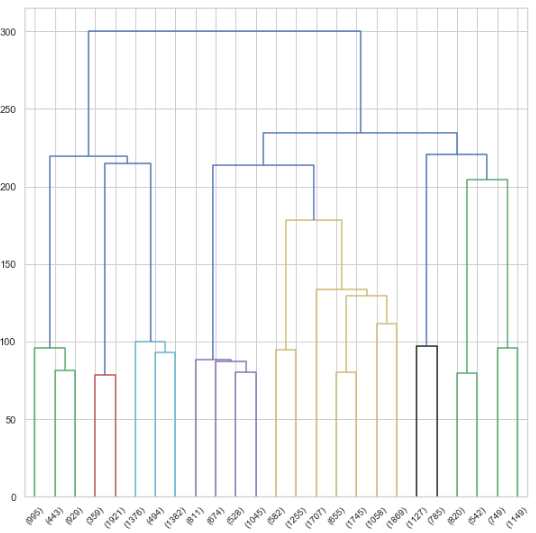


Fig 1.34 Dendogram of hierarchial clustering after Truncating

The cluster grouping linkage based on the dendrogram, 3 or 4 looks good. The further analysis, and based on the dataset had gone for 3 linkage solution based on the hierarchical clustering. From hierarchial clustering we can find there are 7 clusters identified for the data.

In K-means clustering, the inertia is calculated, from that finding plotting the elbow curve and finding the clusters from the elbow curve

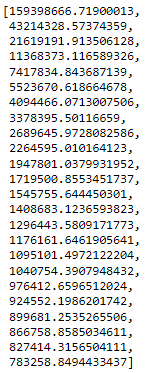


Fig 1.35 Finding inertia from K-means clustering

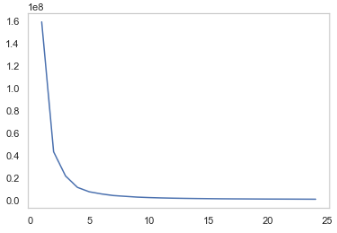


Fig 1.36 Elbow curve for K – means

From the Elbow curve we find that the elbow bends down at 4 cluster linkages are present in the data.

***Silhouette Score and Silhouette width:***



Fig 1.37 Silhouette score



Fig 1.38 Silhouette width

***Silhouette sample value:***



Fig – 1.39 Silhouette Sample value

From this we can clearly identify that we can separate the market zone based on the 4 type of clusters. 4 type of clusters can be

1. High demand & high supply

2. High demand & low supply

3. Low demand & high supply

4. Low demand & low supply

From this we can separate the high performing areas and low performing zones. Based on this we can make promotion, offers, and discounts, make new marketing strategies for making low demand zone to high demand zone and make profits for the company.

*c) Any other business insights:*

**Important business insights are**

* reduce cost,
* Improve the overall organization performance
* customer satisfaction