BUSINESS REPORT

SUPPLY CHAIN PROJECT CAPSTONE

Vishal Srinivasan
PGP DSBA | G3 JULY 2021

TABLE OF CONTENTS

List of Figures

List of Tables

INTRODUCTION

A FMCG company has entered 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 low and where the demand is low, supply is 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 warehouse in entire country.

PROBLEM STATEMENT

Prediction of the shipment of the product each time at the warehouse

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 can 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.

Data dictionary:

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

2 Exploratory Data Analysis (EDA) and Business Implications

2.1 Visual inspection of data

The quick glimpse of the data is shown below

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

5 rows × 24 columns

The dataset has 25000 rows and 24 columns

Types of data present in the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Ware_house_ID                 25000 non-null  object
 1   WH_Manager_ID                 25000 non-null  object
 2   Location_type                 25000 non-null  object
 3   WH_capacity_size              25000 non-null  object
 4   zone                          25000 non-null  object
 5   WH_regional_zone              25000 non-null  object
 6   num_refill_req_l3m            25000 non-null  int64
 7   transport_issue_l1y           25000 non-null  int64
 8   Competitor_in_mkt             25000 non-null  int64
 9   retail_shop_num               25000 non-null  int64
 10  wh_owner_type                 25000 non-null  object
 11  distributor_num               25000 non-null  int64
 12  flood_impacted                25000 non-null  int64
 13  flood_proof                   25000 non-null  int64
 14  electric_supply               25000 non-null  int64
 15  dist_from_hub                 25000 non-null  int64
 16  workers_num                   24010 non-null  float64
 17  wh_est_year                   13119 non-null  float64
 18  storage_issue_reported_l3m    25000 non-null  int64
 19  temp_reg_mach                 25000 non-null  int64
 20  approved_wh_govt_certificate  24092 non-null  object
 21  wh_breakdown_l3m              25000 non-null  int64
 22  govt_check_l3m                25000 non-null  int64
 23  product_wg_ton                25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

A quick glimpse on all the different types of data present in the dataset. The dataset has 8 object type variables, 14 integer type variables and 2 float type variables

Five-point summary of the dataset

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_supply	dist_from_h
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.0000
mean	4.089040	0.773680	3.104200	4985.711560	42.418120	0.098160	0.054640	0.656880	163.5373
std	2.606612	1.199449	1.141663	1052.825252	16.064329	0.297537	0.227281	0.474761	62.7186
min	0.000000	0.000000	0.000000	1821.000000	15.000000	0.000000	0.000000	0.000000	55.0000
25%	2.000000	0.000000	2.000000	4313.000000	29.000000	0.000000	0.000000	0.000000	109.0000
50%	4.000000	0.000000	3.000000	4859.000000	42.000000	0.000000	0.000000	1.000000	164.0000
75%	6.000000	1.000000	4.000000	5500.000000	56.000000	0.000000	0.000000	1.000000	218.0000
max	8.000000	5.000000	12.000000	11008.000000	70.000000	1.000000	1.000000	1.000000	271.0000

The above figure shows the 5-point summary of all the different variables present in the data set

Columns in the dataset

```
Index(['Ware_house_ID', 'WH_Manager_ID', 'Location_type', 'WH_capacity_size',
       'zone', 'WH_regional_zone', 'num_refill_req_l3m', 'transport_issue_l1y',
       'Competitor_in_mkt', 'retail_shop_num', 'wh_owner_type',
       'distributor_num', 'flood_impacted', 'flood_proof', 'electric_supply',
       'dist_from_hub', 'workers_num', 'wh_est_year',
       'storage_issue_reported_l3m', 'temp_reg_mach',
       'approved_wh_govt_certificate', 'wh_breakdown_l3m', 'govt_check_l3m',
       'product_wg_ton'],
      dtype='object')
```

The above figure shows the various columns present in the dataset

Head of the Dataset

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

5 rows × 24 columns

The above figure shows the first 5 rows and columns of the dataset

Tail of the dataset

	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt
24995	WH_124995	EID_74995	Rural	Small	North	Zone 1	3	0	4
24996	WH_124996	EID_74996	Rural	Mid	West	Zone 2	6	0	4
24997	WH_124997	EID_74997	Urban	Large	South	Zone 5	7	0	2
24998	WH_124998	EID_74998	Rural	Small	North	Zone 1	1	0	2
24999	WH_124999	EID_74999	Rural	Mid	West	Zone 4	8	2	4

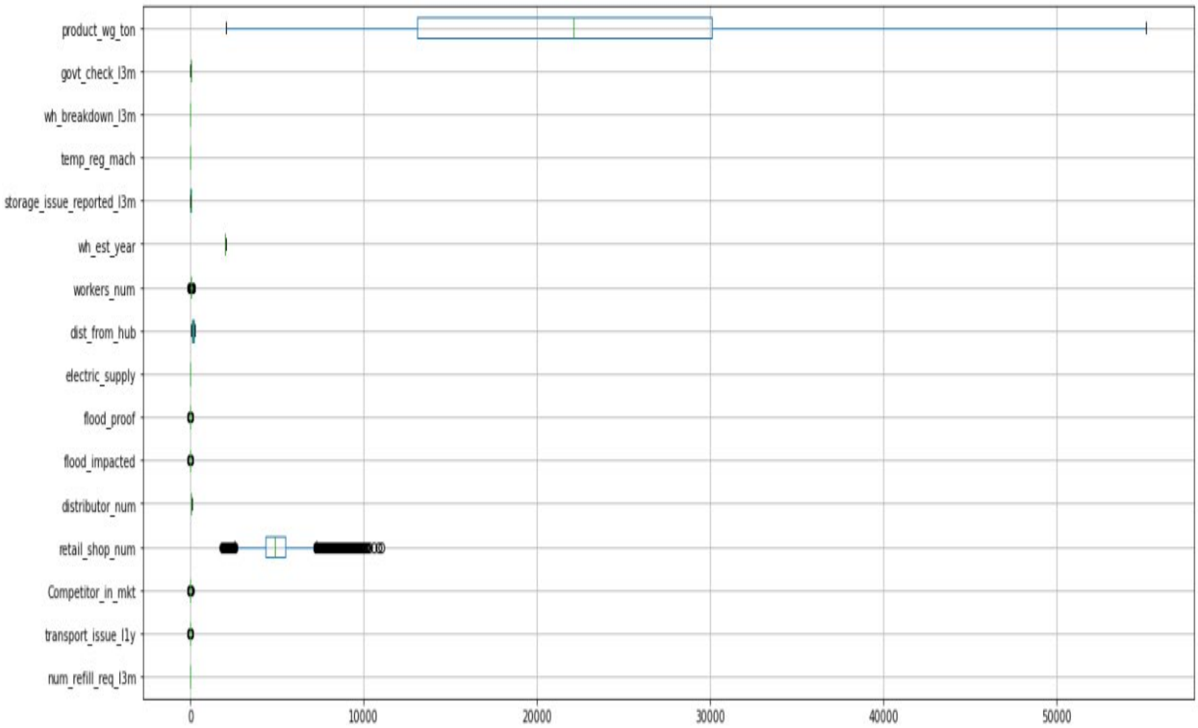
5 rows × 24 columns

The above Figure Shows the last 5 rows and columns
of the Dataset

2.2 Univariate Analysis

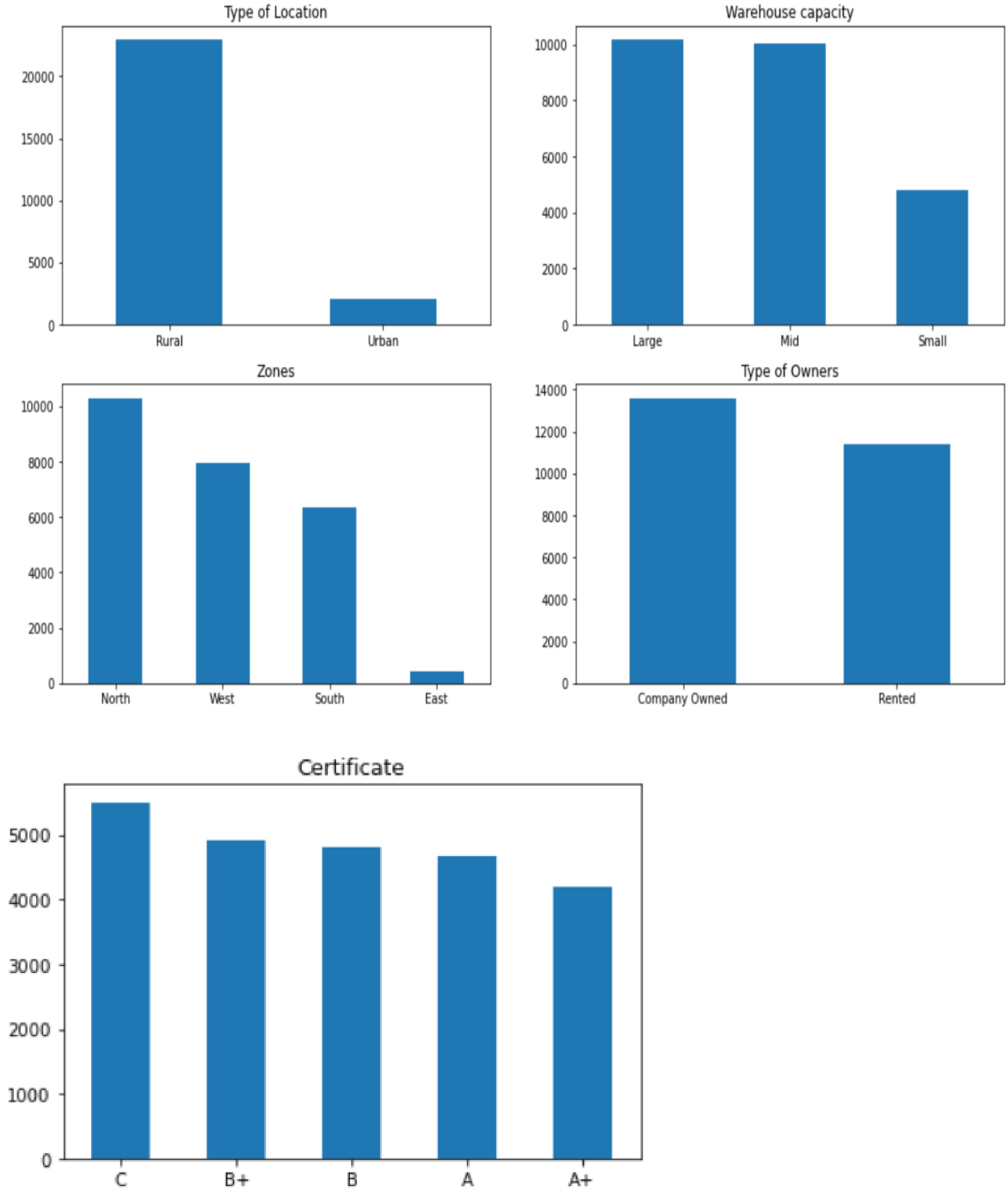
The term **univariate analysis** refers to the analysis of one variable. The purpose of univariate analysis is to understand the distribution of values for a single variable.

Box Plot



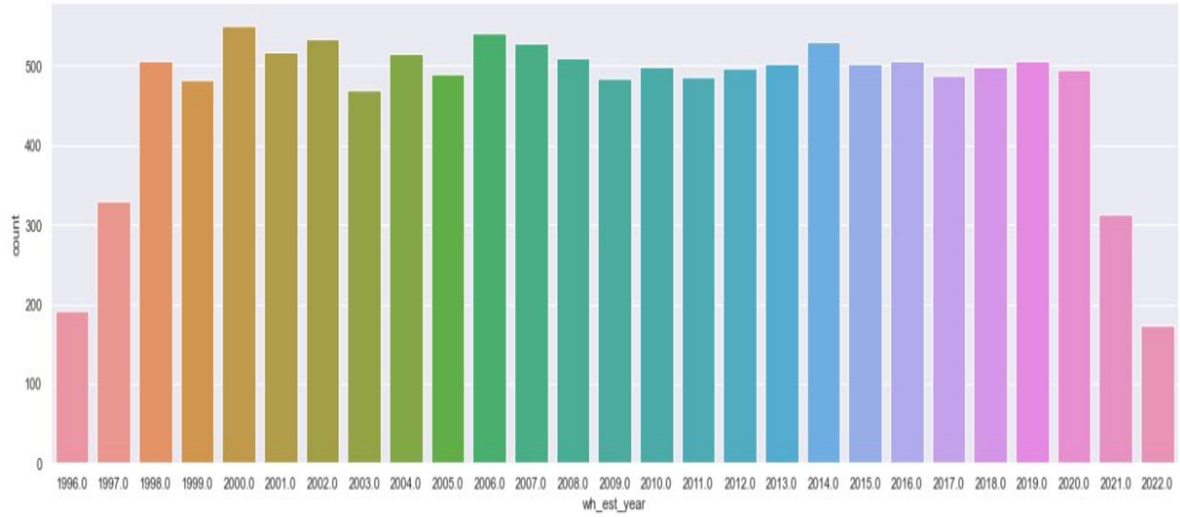
The above figure shows the 5-point summary of the numerical variables present in the datatype in a pictorial manner

BAR PLOT

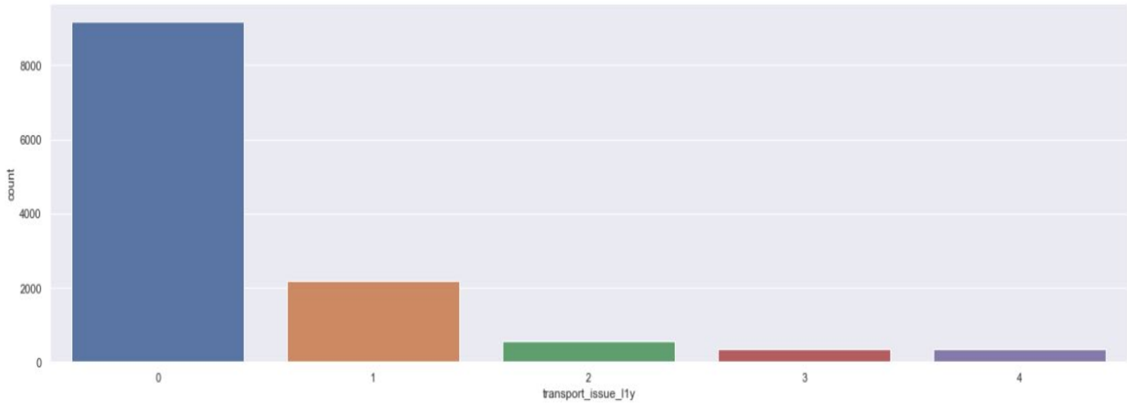


The above figure shows the bar plot for the categorical data present in the Dataset

Out[189]: <seaborn.axisgrid.FacetGrid at 0x226f1271580>



Count Plot for the variable wh_est_year



Count Plot for the variable transport_issue_l1y

Inference

- There are a greater number of Warehouses in the rural area than in the urban area
- The capacity of warehouse of the company is more in the large category followed by the mid category and the small category
- There are greater number of warehouses present in the North Zone followed by the West zone, South zone, and the Least number of Warehouses in the East Zone
- There are a greater number of company owned warehouses than the rented warehouses
- From the Certificate Bar Plot, we find that the warehouses belonging the company has been certified in the "C" certification the greatest number of times

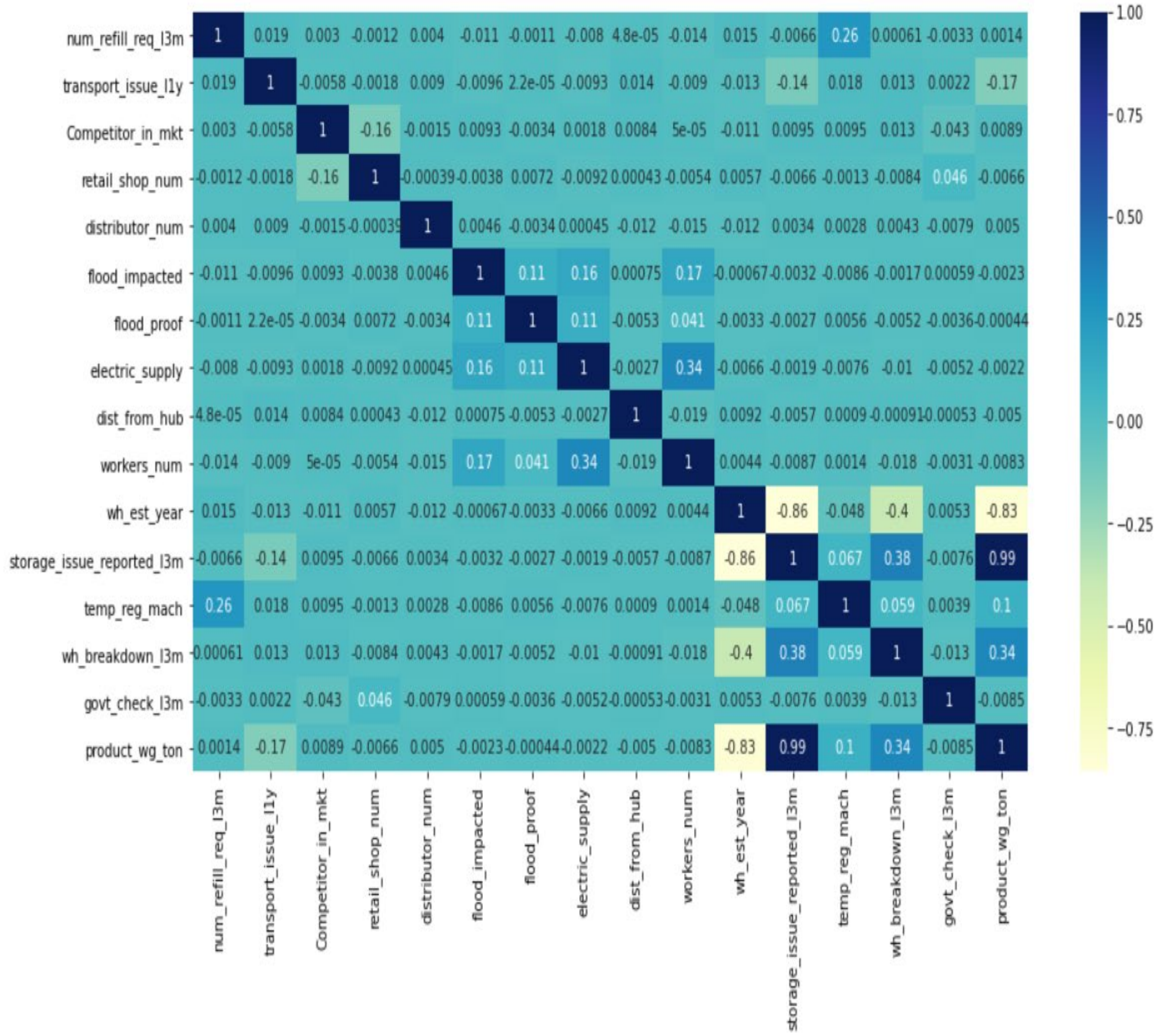
2.3 Bi-variate Analysis:

The Bivariate Analysis is the simplest form of quantitative analysis. It involves the analysis of two variables, and it is used find the empirical relationship between the two variables.

Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable, if we know the value of the other variable

Here we have used Heat map, correlation matrix, Histograms and pair plot for the bivariate analysis

Heat Map

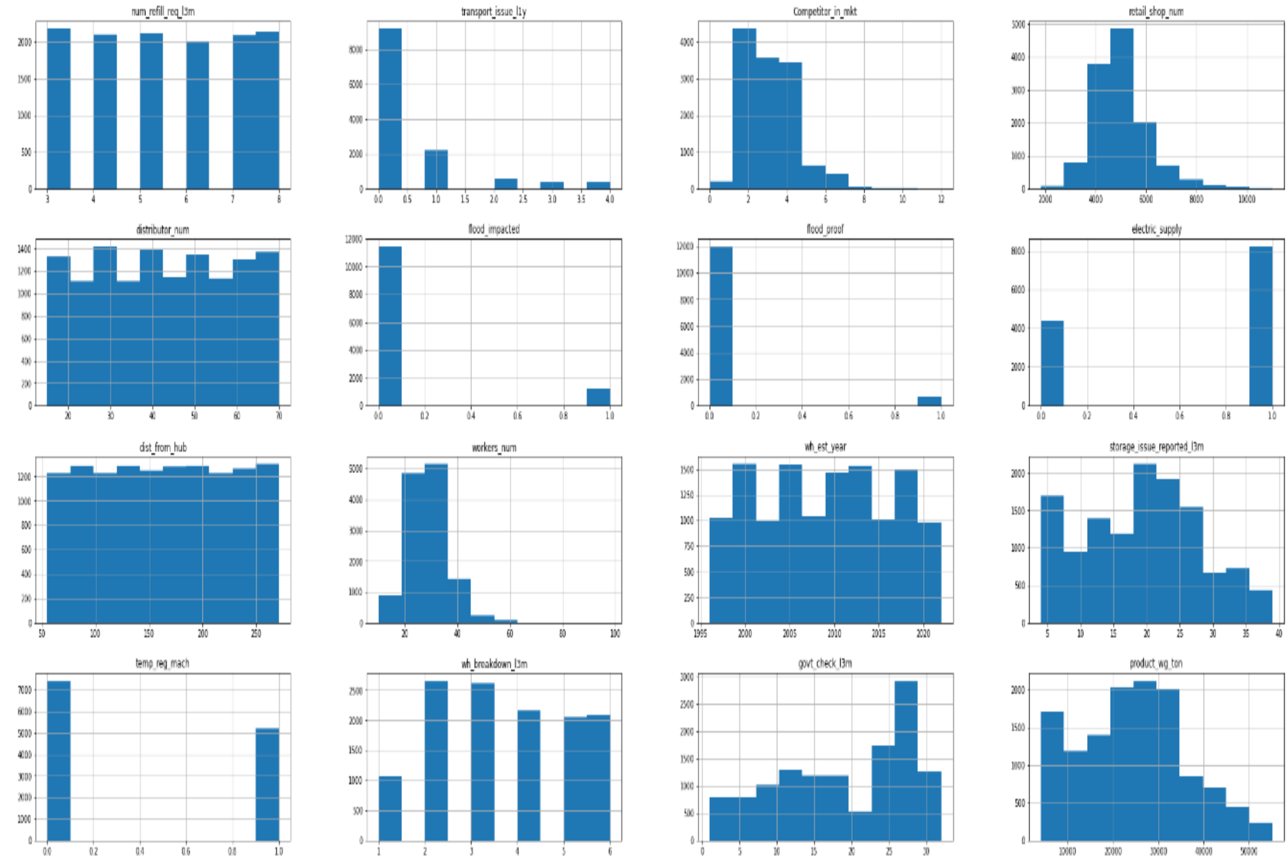


Heat map tells the correlation and collinearity between variables from the above heat map we can observe that there is a heavy correlation and presence of multicollinearity. Multicollinearity is not acceptable in regression

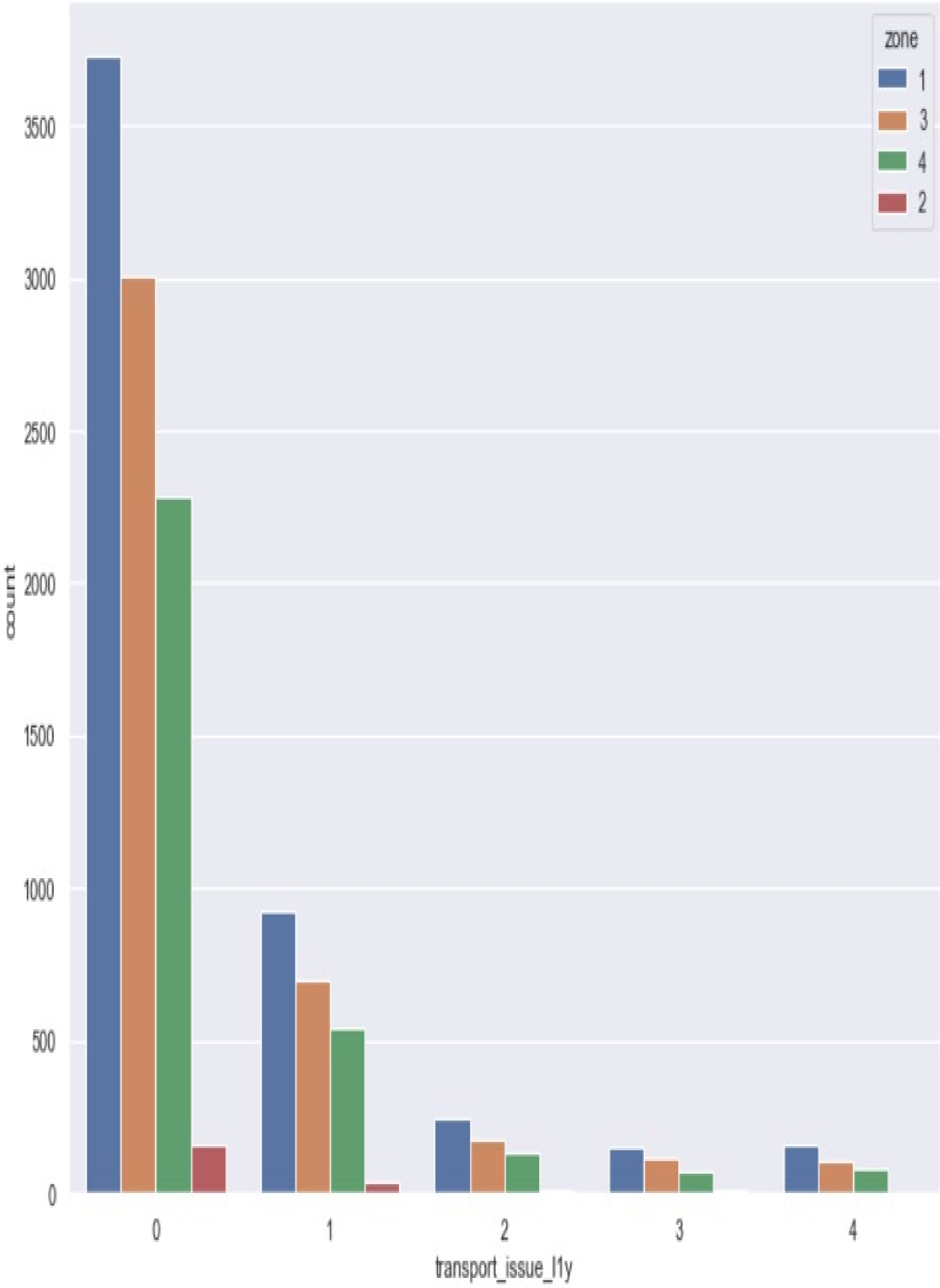
Correlation Matrix

	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num	distributor_num	flood_impacted	flood_proof	electric_
num_refill_req_l3m	1.000000	0.018549	0.002985	-0.001186	0.003995	-0.010548	-0.001123	-0.
transport_issue_l1y	0.018549	1.000000	-0.005826	-0.001826	0.008993	-0.009596	0.000022	-0.
Competitor_in_mkt	0.002985	-0.005826	1.000000	-0.156943	-0.001492	0.009338	-0.003444	0.
retail_shop_num	-0.001186	-0.001826	-0.156943	1.000000	-0.000395	-0.003774	0.007223	-0.
distributor_num	0.003995	0.008993	-0.001492	-0.000395	1.000000	0.004611	-0.003409	0.
flood_impacted	-0.010548	-0.009596	0.009338	-0.003774	0.004611	1.000000	0.107015	0.
flood_proof	-0.001123	0.000022	-0.003444	0.007223	-0.003409	0.107015	1.000000	0.
electric_supply	-0.007959	-0.009299	0.001759	-0.009207	0.000454	0.164815	0.114811	1.
dist_from_hub	0.000048	0.014336	0.008407	0.000429	-0.011838	0.000749	-0.005315	-0.
workers_num	-0.013764	-0.009004	0.000050	-0.005406	-0.014682	0.168425	0.041228	0.
wh_est_year	0.015363	-0.012910	-0.011202	0.005721	-0.012295	-0.000668	-0.003329	-0.
storage_issue_reported_l3m	-0.006602	-0.144327	0.009543	-0.006632	0.003396	-0.003157	-0.002712	-0.
temp_reg_mach	0.260928	0.018207	0.009524	-0.001273	0.002827	-0.008554	0.005636	-0.
wh_breakdown_l3m	0.000608	0.012990	0.012733	-0.008420	0.004286	-0.001744	-0.005151	-0.
govt_check_l3m	-0.003302	0.002190	-0.043455	0.045749	-0.007934	0.000587	-0.003600	-0.
product_wg_ton	0.001415	-0.173992	0.008884	-0.006615	0.004999	-0.002299	-0.000441	-0.

Histogram

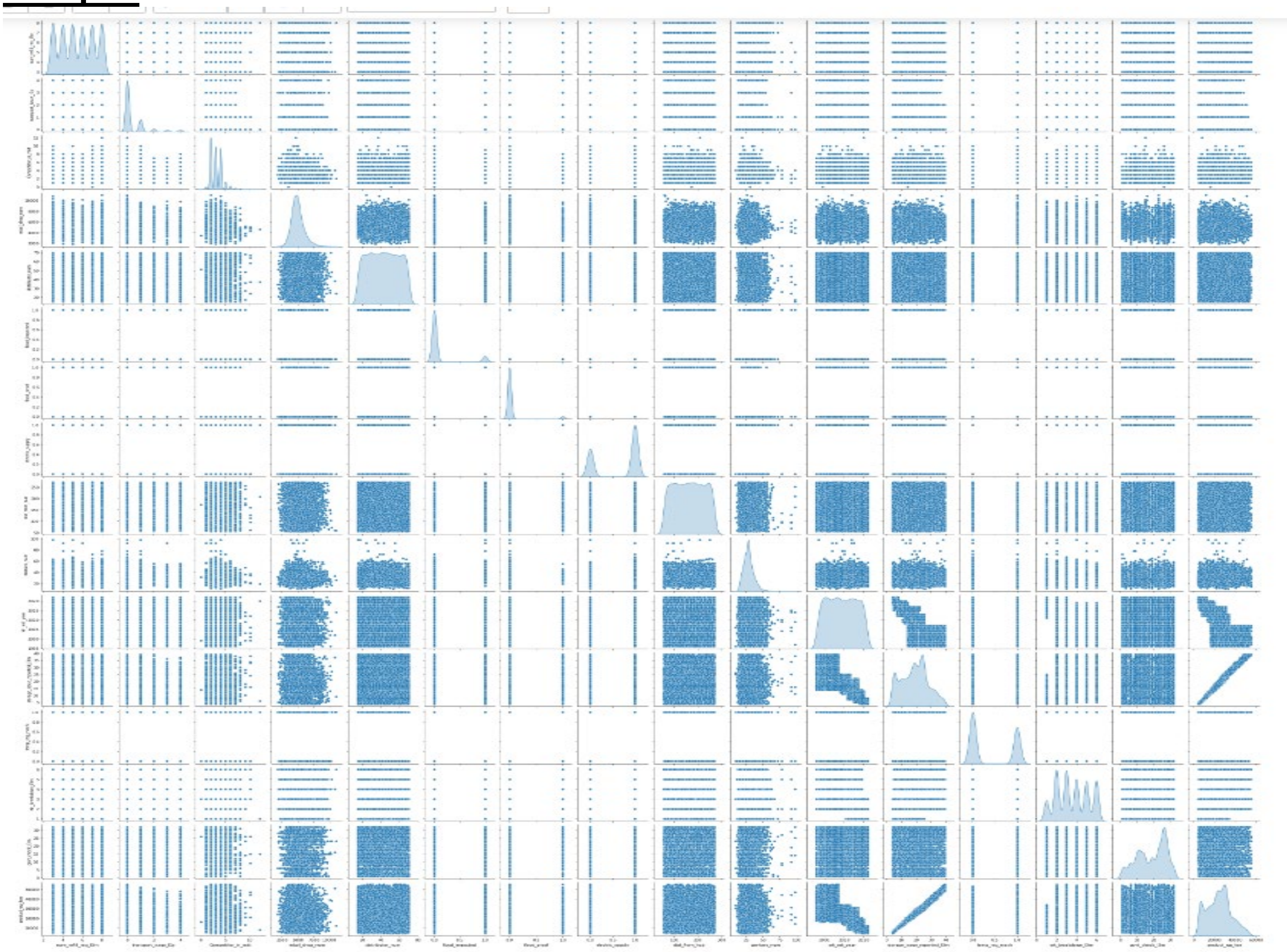


The histogram of all the numerical variables in the dataset



transport_issue_l1y vs Zone

Pair plot



From the above pair plot, we can observe how the data is distributed and relation and patterns between each variable.

3 Data Cleaning and Pre-processing

3.1 Removal of unwanted variables

	Ware_house_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	retail_shop_num
0	WH_100000	Urban	Small	West	Zone 6	3	1	2	4651
1	WH_100001	Rural	Large	North	Zone 5	0	0	4	6217
2	WH_100002	Rural	Mid	South	Zone 2	1	0	4	4306
3	WH_100003	Rural	Mid	North	Zone 3	7	4	2	6000
4	WH_100004	Rural	Large	North	Zone 5	3	1	2	4740
...	...	...	...	...	...	...	...	...	...
24995	WH_124995	Rural	Small	North	Zone 1	3	0	4	5390
24996	WH_124996	Rural	Mid	West	Zone 2	6	0	4	4490
24997	WH_124997	Urban	Large	South	Zone 5	7	0	2	5403
24998	WH_124998	Rural	Small	North	Zone 1	1	0	2	10562
24999	WH_124999	Rural	Mid	West	Zone 4	8	2	4	5664

25000 rows × 23 columns

The WH_Manager_ID is removed as we already have the Ware_house_ID as the index

The Variable storage_issue_reported_l3m is also removed from the dataset as the target variable is what causing the storage issue

3.2 Missing Value treatment

```
Ware_house_ID                    0
WH_Manager_ID                    0
Location_type                    0
WH_capacity_size                 0
zone                             0
WH_regional_zone                 0
num_refill_req_l3m               0
transport_issue_l1y              0
Competitor_in_mkt                0
retail_shop_num                  0
wh_owner_type                    0
distributor_num                  0
flood_impacted                   0
flood_proof                      0
electric_supply                  0
dist_from_hub                    0
workers_num                    990
wh_est_year                  11881
storage_issue_reported_l3m       0
temp_reg_mach                    0
approved_wh_govt_certificate   908
wh_breakdown_l3m                 0
govt_check_l3m                   0
product_wg_ton                   0
dtype: int64
```

The above figure shows us the total number of missing values present in the dataset

The variable workers_num has 990 missing variables and it can be replaced by the mean imputer function as the distribution is normally distributed

The variable wh_est_year has 11881 missing variables, and all the missing values are removed. From the Pair plot we found out that this is one of the variable which has the most relation with the target variable, so instead of dropping the whole variable we have just dropped the missing variables from the variable

	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	re
4	100004	EID_50004	1	3	1	5	3	1	2	
5	100005	EID_50005	1	1	3	1	8	0	2	
6	100006	EID_50006	1	3	3	6	8	0	4	
8	100008	EID_50008	1	1	4	6	8	1	4	
10	100010	EID_50010	1	3	1	6	7	1	3	

5 rows × 24 columns

The above figure shows the new data frame after the wh_est_year variable has been dropped

The approved_wh_govt_certificate variable has 908 NA variables, and all these variables are off the years 2021,2022, 2023.From this we find that newly established warehouses have not been certified by the government

3.3 Outlier treatment

```
num_refill_req_l3m         -0.075217
transport_issue_l1y         1.610907
Competitor_in_mkt           0.978456
retail_shop_num             0.908302
distributor_num             0.015213
flood_impacted              2.701327
flood_proof                 3.919343
electric_supply            -0.660933
dist_from_hub              -0.005999
workers_num                 1.059911
wh_est_year                 0.012417
storage_issue_reported_l3m  0.113345
temp_reg_mach               0.855960
wh_breakdown_l3m           -0.068026
govt_check_l3m             -0.363262
product_wg_ton              0.331631
dtype: float64
```

From the above figure we find the skewness is between -5 to+5 so no outlier treatment is required

4 Model building

STEP 1: All the object variables are converted to Categorical codes

	Ware_house_ID	WH_Manager_ID	Location_type	WH_capacity_size	zone	WH_regional_zone	num_refill_req_l3m	transport_issue_l1y	Competitor_in_mkt	re
4	100004	EID_50004	1	3	1	5	3	1	2	
5	100005	EID_50005	1	1	3	1	8	0	2	
6	100006	EID_50006	1	3	3	6	8	0	4	
8	100008	EID_50008	1	1	4	6	8	1	4	
10	100010	EID_50010	1	3	1	6	7	1	3	

5 rows × 24 columns

The Data set after all the object variables are converted to categorical codes

STEP 2: Constructed a X and y matrix for modelling

```
Index(['Ware_house_ID', 'Location_type', 'WH_capacity_size', 'zone',
       'WH_regional_zone', 'num_refill_req_l3m', 'transport_issue_l1y',
       'Competitor_in_mkt', 'retail_shop_num', 'wh_owner_type',
       'distributor_num', 'flood_impacted', 'flood_proof', 'electric_supply',
       'dist_from_hub', 'workers_num', 'wh_est_year', 'temp_reg_mach',
       'approved_wh_govt_certificate', 'wh_breakdown_l3m', 'govt_check_l3m'],
      dtype='object')
```

The X and y matrix was constructed for model building and the Target Variable product_wg_ton is taken in the y matrix

STEP 3: The dataset is divided into test and train set in a 70:30 ratio. This is done by invoking the sklearn. model_selection import train_test_split.

STEP 4: Model Building

The Following Models were Built
1.Linear Regression Model
2.CART Regressor Model
3.Random Forest Regressor Model
4.AdaBoost Regressor Model
5.Lasso Regressor Model
6.Ridge Regressor Model

4.1 Linear Regression Model

Linear regression is a Linear approach for modelling the relationship between a scalar response and one or more explanatory variables (Dependent and independent variables). The case of one explanatory variable is called as simple linear regression, for more than one, the process is called **multiple linear regression**

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T}\boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Coefficient of Independent Variables

	Coefficient
Ware_house_ID	14.262565
Location_type	34.084890
WH_capacity_size	-12.258380
zone	34.332343
WH_regional_zone	34.492109
num_refill_req_l3m	6.063320
transport_issue_l1y	-1026.934133
Competitor_in_mkt	3.719068
retail_shop_num	86.756451
wh_owner_type	21.918797
distributor_num	54.434782
flood_impacted	-0.133047
flood_proof	29.802403
electric_supply	21.892348
dist_from_hub	-44.803076
workers_num	29.482511
wh_est_year	-9216.431789
temp_reg_mach	664.429022
approved_wh_govt_certificate	-674.340622
wh_breakdown_l3m	344.053214
govt_check_l3m	-92.083553

INTERCEPT

```
24432.312019882425
```

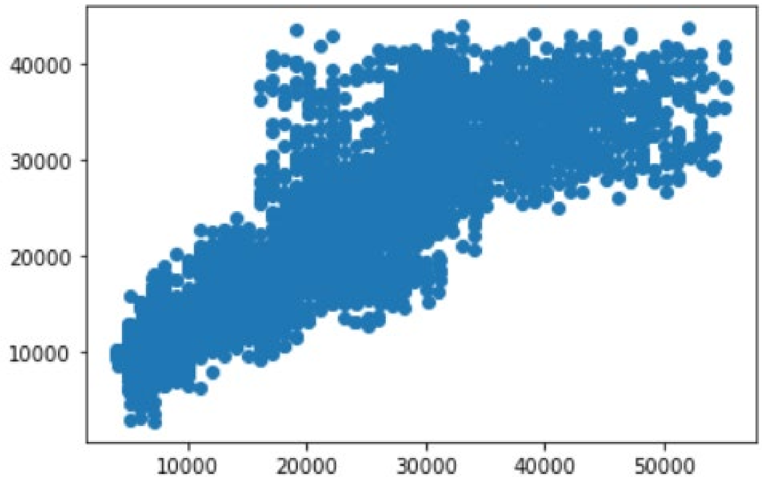
The Intercept of the Model was calculated

Now we have coefficient and intercept to form an equation but how to check the performance of the equation. We can check the performance of the equation by R Square or RMSE

Evaluation Metrics of Linear Regression Model

	TEST	TRAIN
MAE	5052.314	4988.67
MSE	43900552.56	42205437.39
RMSE	6625.75	6496.58
R2	0.676	0.681

Scatter Plot for Actual Value and Predicted Value



4.2 CART Regressor Model

Evaluation Metrics of CART Regressor Model

	TEST	TRAIN
MAE	5095.162	5008.395
MSE	42579164.7	40322828.91
RMSE	6525.27	6350.02
R2	0.6864	0.6958

After Doing the Grid Search CV the RMSE value came down to 6200.957

4.3 Random Forest Regressor Model

Evaluation Metrics of Random Forest Regressor Model

	TEST	TRAIN
MAE	4719.21	1709.5
MSE	3843201.07	4923235.76
RMSE	6199.356	2218.83
R2	0.7169	0.9628

After Doing the Grid Search CV the RMSE value is 6201.297, as the RMSE value after Grid Search CV is higher, the model performs good even without tuning

4.4 AdaBoost Regressor Model

Evaluation Metrics of AdaBoost Regressor Model

	TEST	TRAIN
MAE	4705.18	85.756
MSE	40485522.3	115082.32
RMSE	6362.823	339.237
R2	0.7018	0.999132

After Doing the Grid Search CV the RMSE value is 6408.270, as the RMSE value after Grid Search CV is higher, the model performs good even without tuning

4.5 Lasso Regressor Model

Evaluation Metrics of Lasso Regressor Model

	TEST	TRAIN
MAE	9102.75	8986.66
MSE	125328989.1	121216071.7
RMSE	11195.043	11009.81
R2	0.07694	0.085

After Doing the Grid Search CV the RMSE value came down to 6623.298 and

R2 value went up to 0.68

4.6 Ridge Regressor Model

Evaluation Metrics of Ridge Regressor Model

	TEST	TRAIN
MAE	5056.9	4994.42
MSE	43928981.79	42217135.75
RMSE	6627.894	6497.47
R2	0.67646	0.6815

After Doing the Grid Search CV the RMSE value came down to 6625.885 and R2 value went up to 0.679

4.7 Hyper Parameter Tuning

Hyper-parameter tuning refers to the process of find hyper-parameters that yield the best result. This, of course, sounds a lot easier than it actually is. Finding the best hyper-parameters can be an elusive art, especially given that it depends largely on your training and testing data.

As your data evolves, the hyper-parameters that were once high performing may no longer perform well. Keeping track of the success of your model is critical to ensure it grows with the data.

One way to tune your hyper-parameters is to use a grid search. This is probably the simplest method as well as the crudest. In a grid search, you try a grid of hyper-parameters and evaluate the performance of each combination of hyper-parameters.

Model Name	Before Tuning	After Tuning
Linear Regression	6625.75	6625.75
CART	6525.27	6200.957
Random Forest	6199.356	6201.297
Adaptive Boost	6362.823	6408.27
Lasso	11195.043	6623.298
Ridge	6627.894	6625.885

The RMSE values of all the models used after and before tuning are listed above

5. Model validation

Regression model evaluation metrics

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.

- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.

- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.

- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

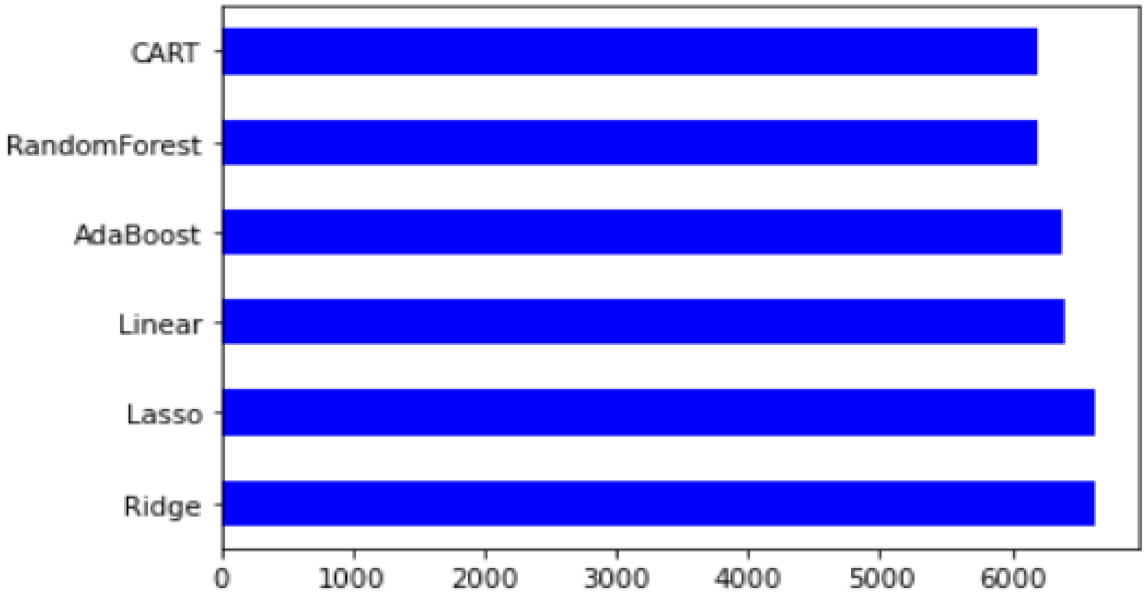
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of* y

$\bar{y}$ $-$ *mean value of* y



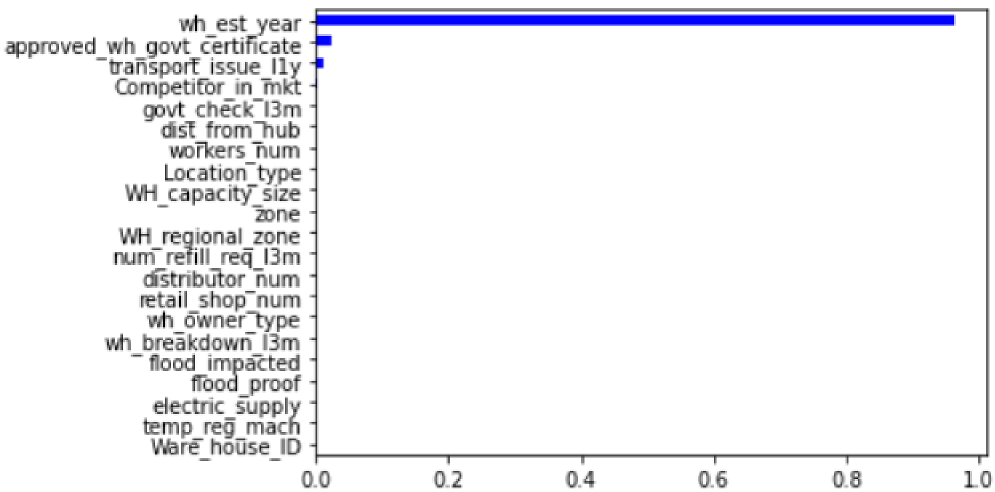
Model Comparison

The above plot shows the model performance based on the RMSE value

From the above Plot we come to know that the Random Forest Regressor Model is the optimum model to be used for Prediction as per the RMSE value

<u>6. Final interpretation / recommendation</u>

The Final interpretation is based on feature Importance
From the most optimum model.i.e., the Random Forest Model, we find the most important features which influences the target variable



Feature Importance
From the above plot we find that the wh_est_year, approved_wh_govt_certificate and the transport_issue_1ly are the most important features

Insights

- It is observed that the greatest number of warehouses are located at Zone-6 an d this further indicates that there is higher demand in that Zone

- The availability of generator in the warehouse doesn't affect much

- About 54.3% of the warehouses are company owned and the rest 45.7% are re nted

- There are a greater number of Warehouses in the rural area than in the urban area

- The capacity of warehouse of the company is more in the large category followed by the mid category and the small category

- There are greater number of warehouses present in the North Zone followed by the West zone, South zone, and the Least number of Warehouses in the East Zone

- From the Certificate Bar Plot, we find that the warehouses belonging the company has been certified in the "C" certification the greatest number of times

- It is clear from the observations that the variable"storage_issue_reported_l3m" plays an important role in suggesting the company about the quantity to be shipped

- North Zone has the highest significance than the other zones

Recommendations

- The oldest warehouse which is established in the year 1996 and it needs to be maintained properly with all the prerequisites before the product is shipped to those warehouses

- Some frequent repairs to be done on older warehouses so that the storage issues faced by those warehouses can be prevented

- The transport issues faced by the warehouses also influences the quantity to be shipped to the warehouses, so those issues should be addressed by the company so that the quantity stored in the warehouses can be optimized and supply chain issues can be mitigated

- It is found that from the observations the warehouses with A+ certification from the government performs well in the business than the warehouses with other certifications and the company should strive to get A+ certifications so that the company has a greater number of warehouses that performs well

- The areas where there is more number of competitors should be addressed and more number of warehouses should be established in those areas

- The distance from the hub impacts the business as it induces the transport issue and this issue to be addressed by the company

- The temperature of machines in the warehouses should be monitored carefully as it may give rise to storage issues

- There are 5896 warehouses that are under performing and those warehouses should be reassessed and preventive efforts to be made by the company like increasing number of workers working in the warehouses

- The Location of the warehouse plays an important role in supply and demand of the product and those locations where there is low demands less quantity of product can be shipped and increasing the product quantity in areas of higher demand

- In areas of high demand the retail shops should be increased so that there is enough product to be sold