CAPSTONE PROJECT

FINAL REPORT

(BUSINESS REPORT)

BY

DANIEL KAMAL RAJ KOPPETI

BATCH: OCT A – GROUP 8

DATE – 09-10-2022

Table of Contents

List of Figures

List of Tables

1. Introduction

Problem statement

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

Need of the study

Due to the miss match in the demand and supply the company is suffering a loss in the inventory cost and the company would want to solve this. This can be solved by performing a good analysis and then build a model using historical data provided that will determine an optimum weight of the product to be shipped each time to the warehouse so that there would be no loss in inventory costs or operational costs and also there is no mismatch between demand and supply.

And also, to analysis the demand pattern in different pockets of the country so management can drive the advertisement campaign particular in those pockets so that the company gets value based on this analysis.

2. EDA and Business Implication

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ware_house_ID	25000	25000	WH_100000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_Manager_ID	25000	25000	EID_50000	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Location_type	25000	2	Rural	22957	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_capacity_size	25000	3	Large	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zone	25000	4	North	10278	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WH_regional_zone	25000	6	Zone 6	8339	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num_refill_req_l3m	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
transport_issue_l1y	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
Competitor_in_mkt	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
retail_shop_num	25000.0	NaN	NaN	NaN	4985.71156	1052.825252	1821.0	4313.0	4859.0	5500.0	11008.0
wh_owner_type	25000	2	Company Owned	13578	NaN	NaN	NaN	NaN	NaN	NaN	NaN
distributor_num	25000.0	NaN	NaN	NaN	42.41812	16.064329	15.0	29.0	42.0	56.0	70.0
flood_impacted	25000.0	NaN	NaN	NaN	0.09816	0.297537	0.0	0.0	0.0	0.0	1.0
flood_proof	25000.0	NaN	NaN	NaN	0.05464	0.227281	0.0	0.0	0.0	0.0	1.0
electric_supply	25000.0	NaN	NaN	NaN	0.65688	0.474761	0.0	0.0	1.0	1.0	1.0
dist_from_hub	25000.0	NaN	NaN	NaN	163.53732	62.718609	55.0	109.0	164.0	218.0	271.0
workers_num	24010.0	NaN	NaN	NaN	28.944398	7.872534	10.0	24.0	28.0	33.0	98.0
wh_est_year	13119.0	NaN	NaN	NaN	2009.383185	7.52823	1996.0	2003.0	2009.0	2016.0	2023.0
storage_issue_reported_l3m	25000.0	NaN	NaN	NaN	17.13044	9.161108	0.0	10.0	18.0	24.0	39.0
temp_reg_mach	25000.0	NaN	NaN	NaN	0.30328	0.459684	0.0	0.0	0.0	1.0	1.0
approved_wh_govt_certificate	24092	5	C	5501	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wh_breakdown_l3m	25000.0	NaN	NaN	NaN	3.48204	1.690335	0.0	2.0	3.0	5.0	6.0
govt_check_l3m	25000.0	NaN	NaN	NaN	18.81228	8.632382	1.0	11.0	21.0	26.0	32.0
product_wg_ton	25000.0	NaN	NaN	NaN	22102.63292	11607.755077	2065.0	13059.0	22101.0	30103.0	55151.0

Table 1 : Summary of the dataset

- The data set consists of 25000 rows and 24 columns.
- The column product_wg_ton has a mean value of 22102.63 and a median value of 22101. The mean and median values are almost equal which suggests normal distribution. The min and max values are 2065 and 55151 respectively.
- The column dist_from_hub has a mean value of 163.53 and median value of 164. The min and max values are 55 and 271 respectively.
- The columns Ware_house_ID and WH_Manager_ID seems to be the unique identifiers.
- The other descriptive statistics of the columns can be observed from the above table.

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

Table 2 : Data Dictionary

```
 #   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
```

Figure 1 : Information about the dataset.

- The columns Ware_house_ID, WH_Manager_ID, Location_Type, WH_capacity_size, zone, WH_regional_zone, wh_owner_type, approved_wh_govt_certificate is of object datatype and all the remaining columns are of numeric datatype.
- workers_num, wh_est_year, approved_wh_govt_certificate have null values.
- Renaming of columns is not required.

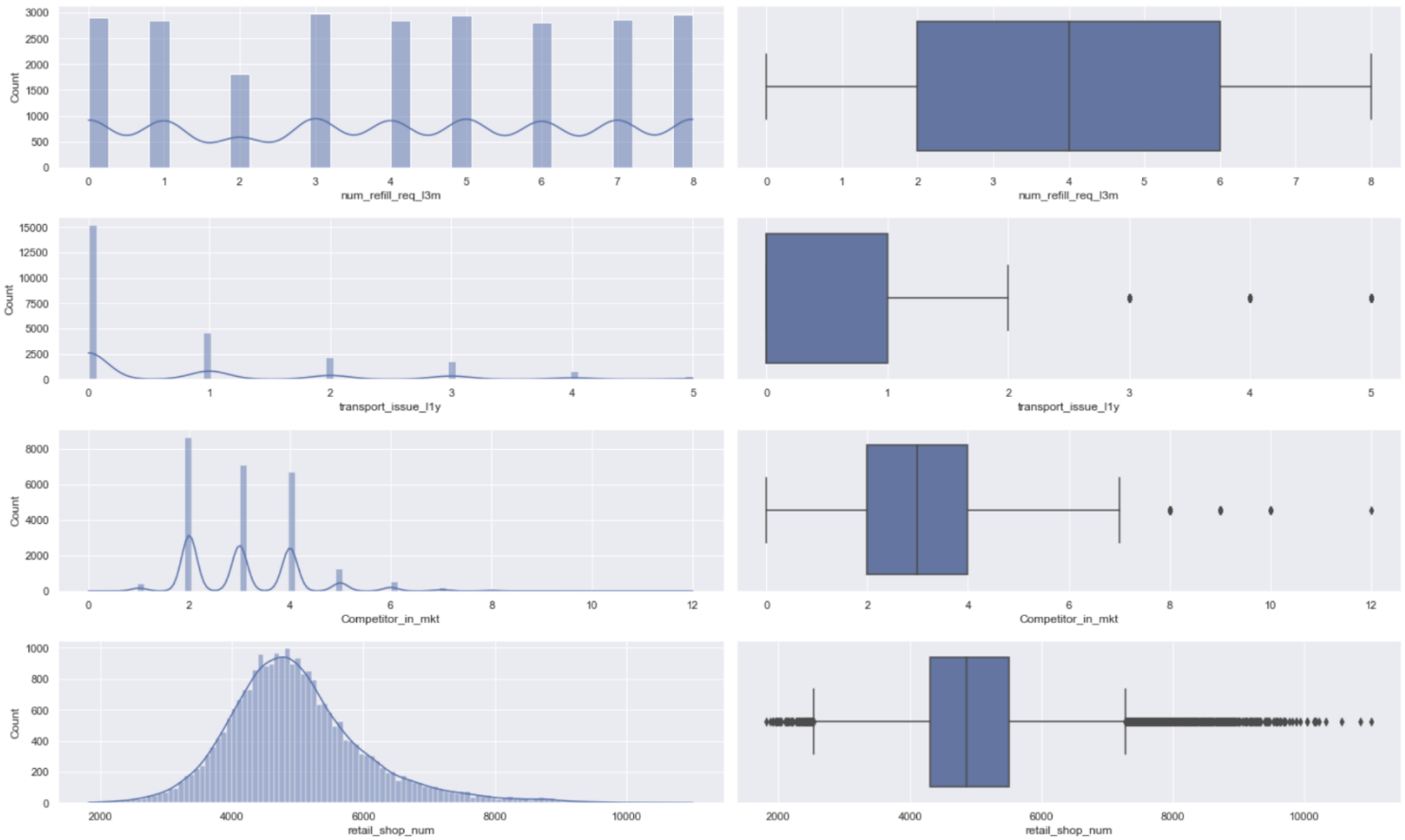
a) Univariate analysis

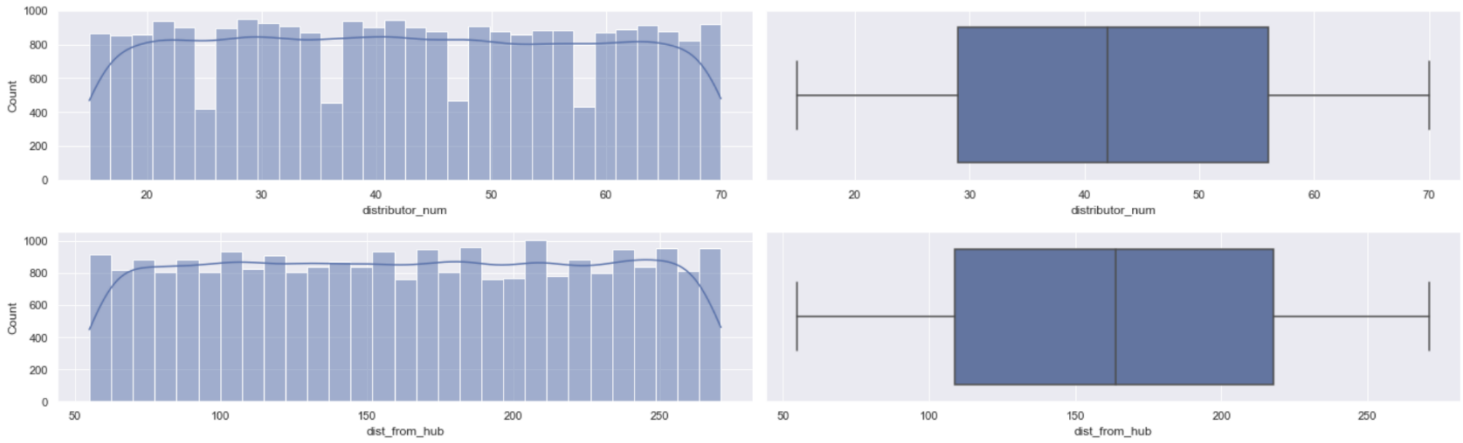
The are some columns which are categorical in nature but are of integer type. Hence converting them into string type for further analysis.

Those columns are **'flood_impacted','flood_proof','electric_supply','temp_reg_mach'.** The column wh_est_year is not considered for analysis as it is year.

Univariate Analysis

Columns - num_refill_req_l3m, transport_issue_l1y, Competitor_in_mkt, retail_shop_num, distributor_num, dist_from_hub





Columns - workers_num, storage_issue_reported_l3m, wh_breakdown_l3m, govt_check_l3m, product_wg_ton

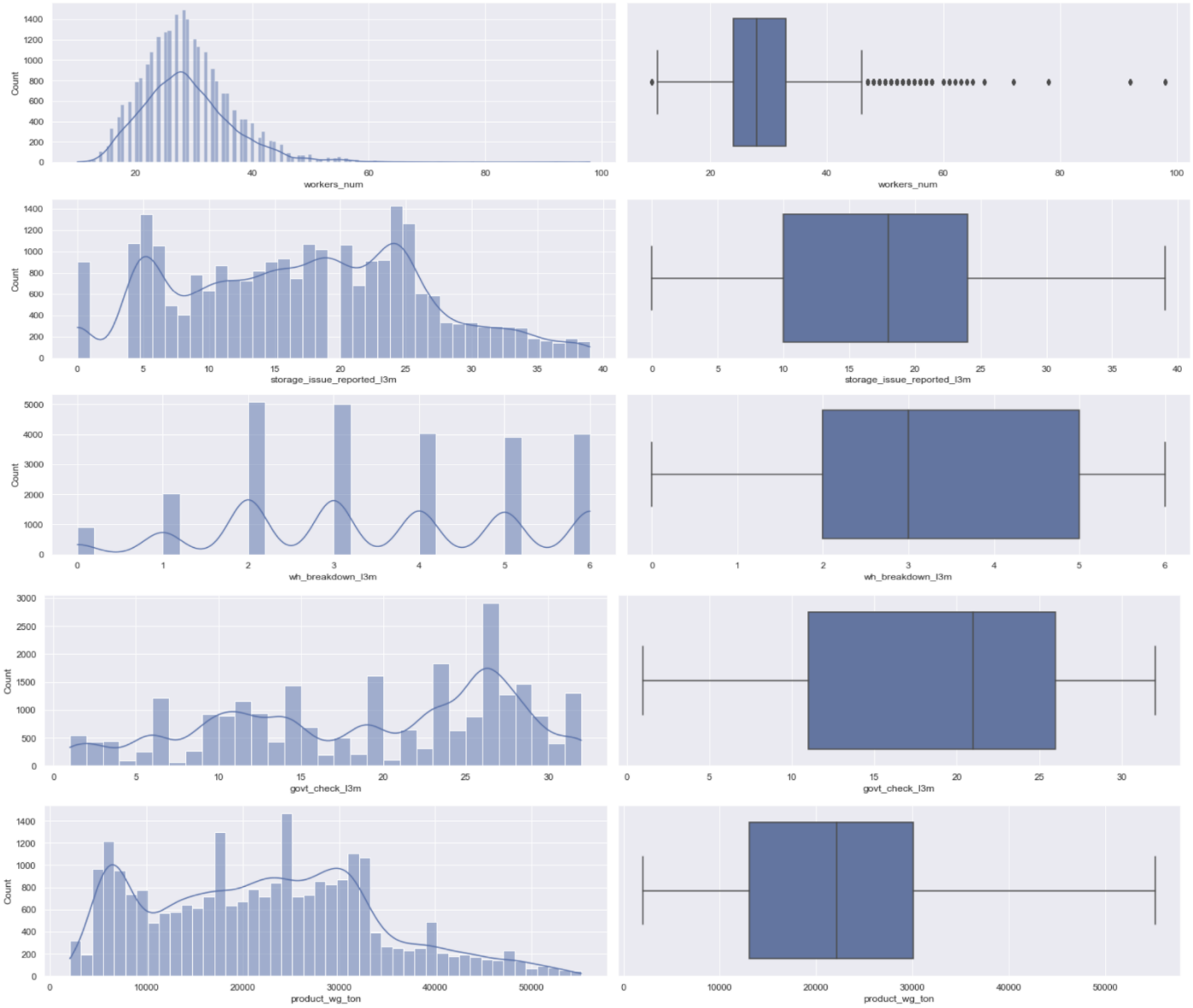
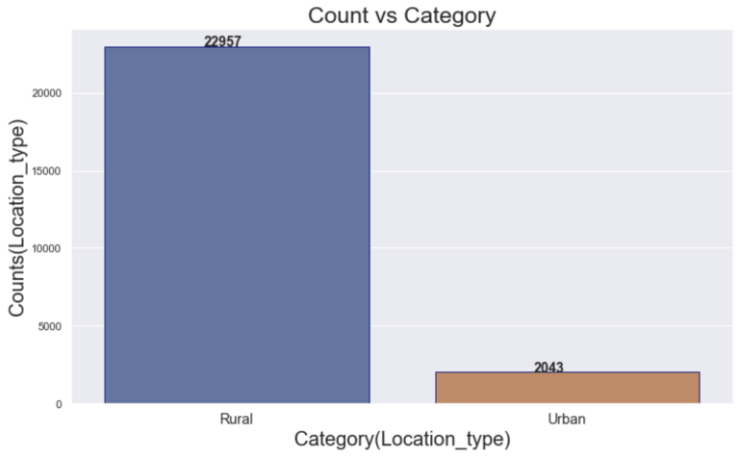
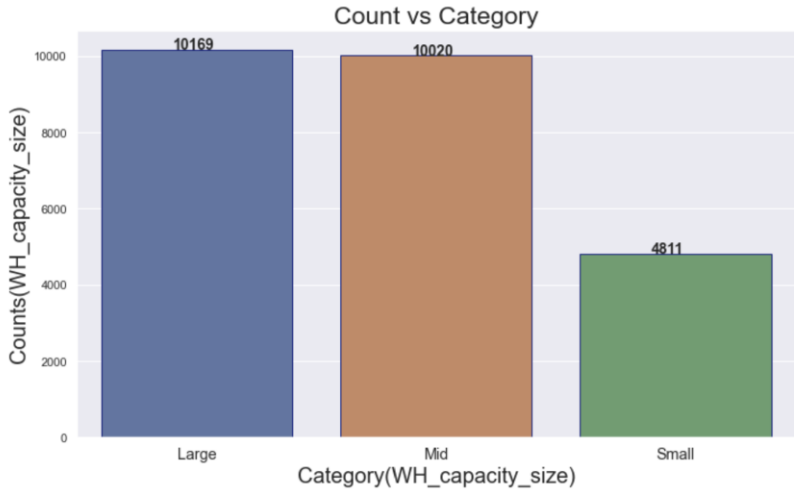


Figure 2 : Histograms and Boxplots of Continuous Variables

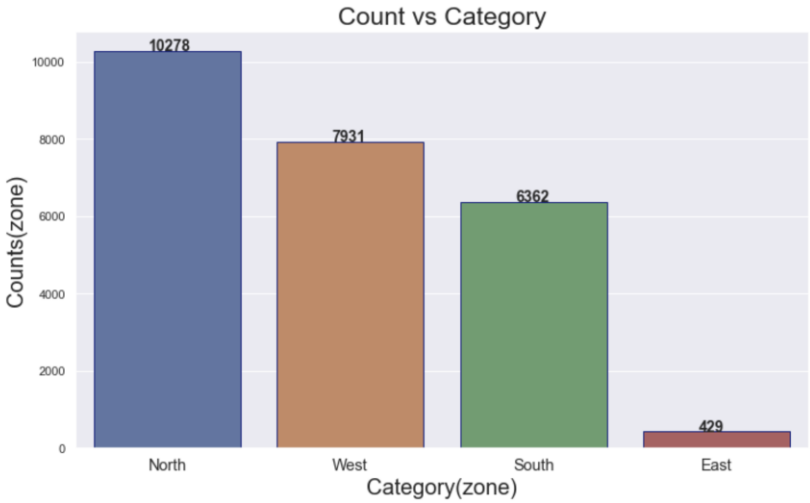
- The columns transport_issue_l1y, Competitor_in_mkt, retail_shop_num,workers_num have outliers



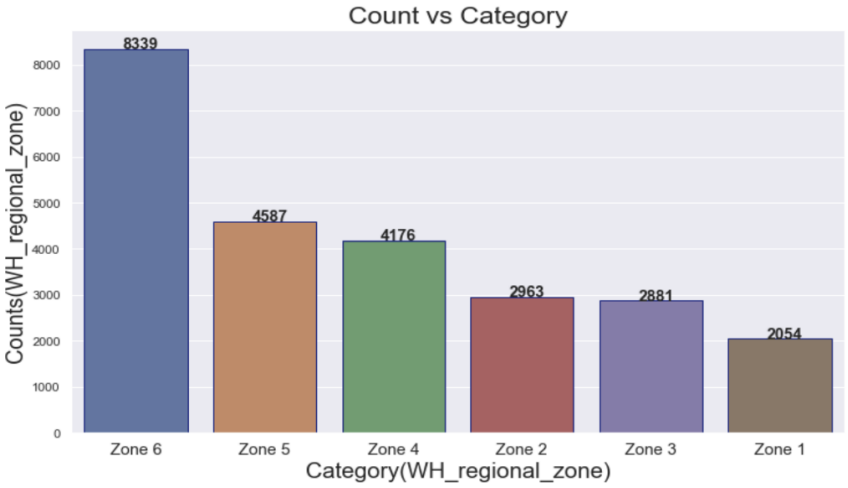
- The column Location_type has 2 categories. They are Rural and Urban and their counts are 22957 and 2043 respectively.



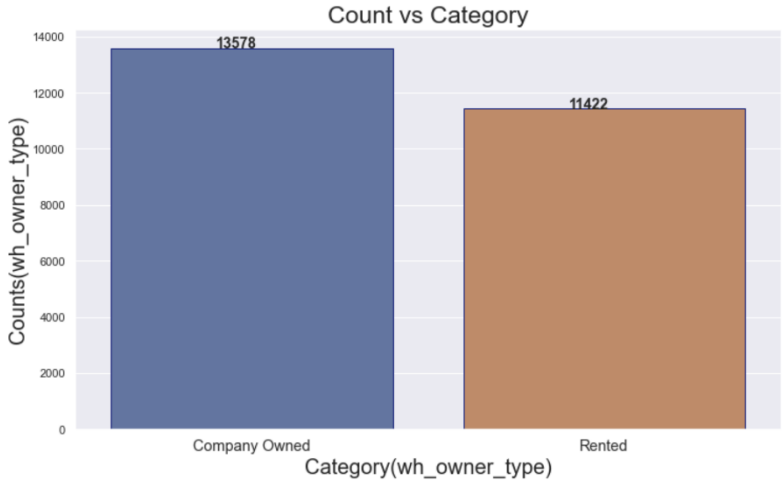
- The column WH_capacity_size has 3 categories. They are Large, Mid and Small and their counts are 10169, 10020 and 4811 respectively.



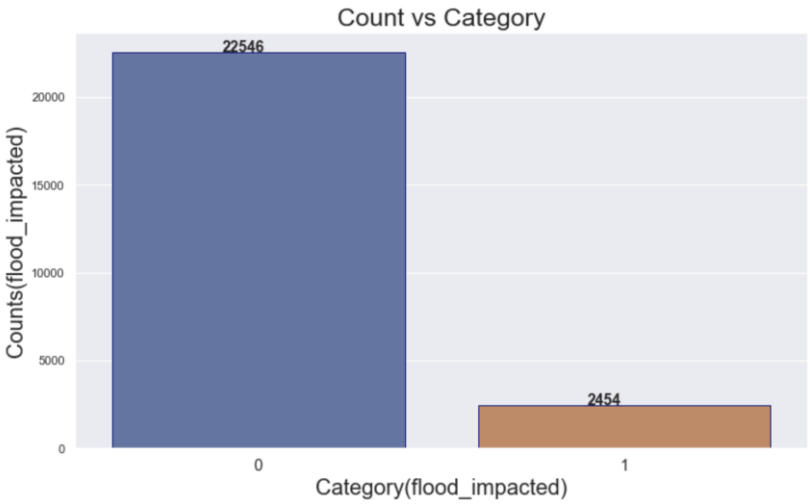
- The column zone has 4 categories. They are North, West, South and East and their counts are 10278, 7931 6362 and 429 respectively.



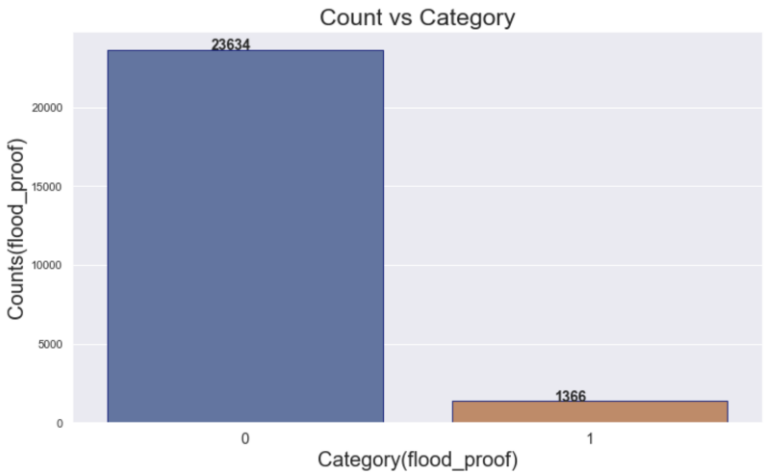
- The column WH_regional_zone has 6 categories. They are Zone6, Zone5, Zone4, Zone3, Zone2 and Zone1 and their counts are 8339, 4587, 4176, 2963, 2881 and 2054 respectively.



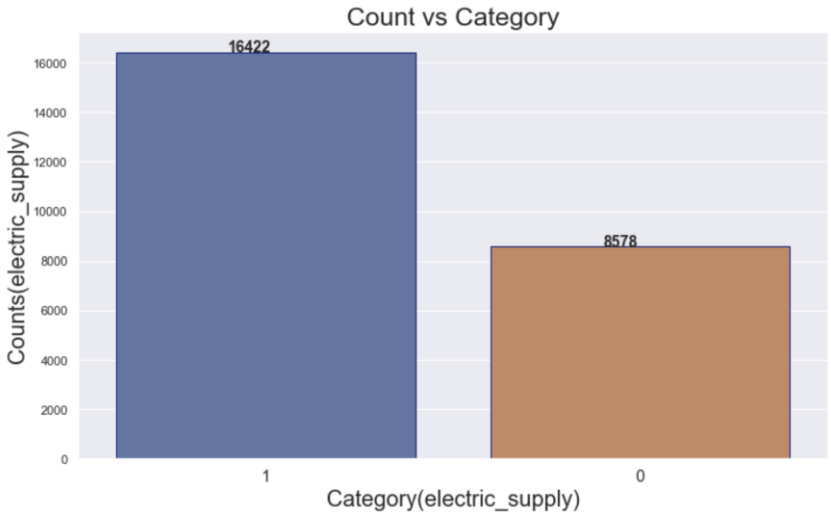
- The column wh_owner_type has 2 categories. They are Company owned and Rented and their counts are 13578 and 11422 respectively.



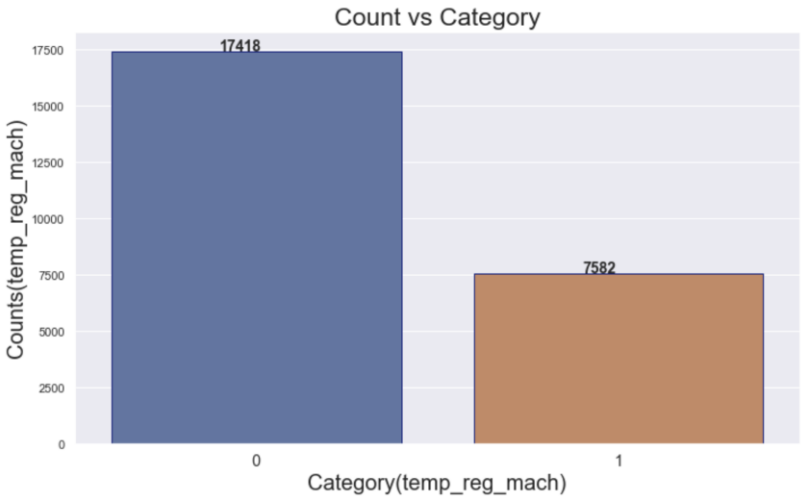
- The column flood_impacted has 2 categories. They are 0 and 1 and their counts are 22546 and 2454 respectively.



- The column flood_proof has 2 categories. They are 0 and 1 and their counts are 23634 and 1366 respectively.



- The column electric_supply has 2 categories. They are 1 and 0 and their counts are 16422 and 8578 respectively.



- The column temp_reg_match has 2 categories. They are 0 and 1 and their counts are 17418 and 7582 respectively.

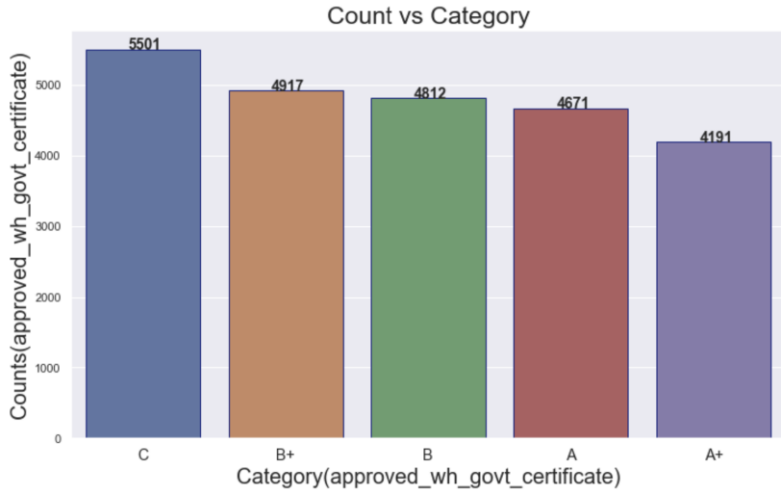


Figure 3 : Count plots of Categorical Variables

- The column approved_wh_govt_certificate has 6 categories. They are C, B+, B, A and A+ and their counts are 5501, 4917, 4812, 4671 and 4191 respectively.

Bivariate Analysis

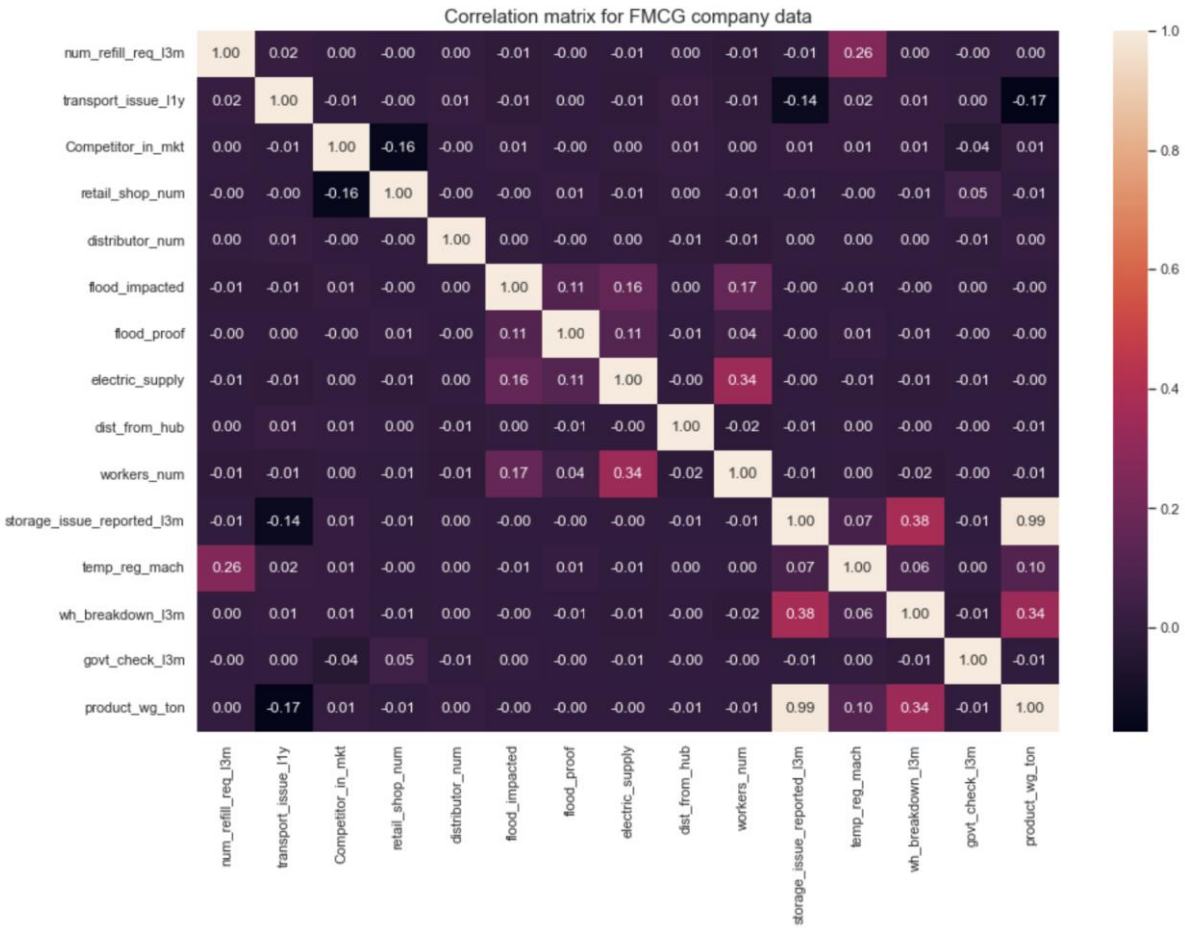
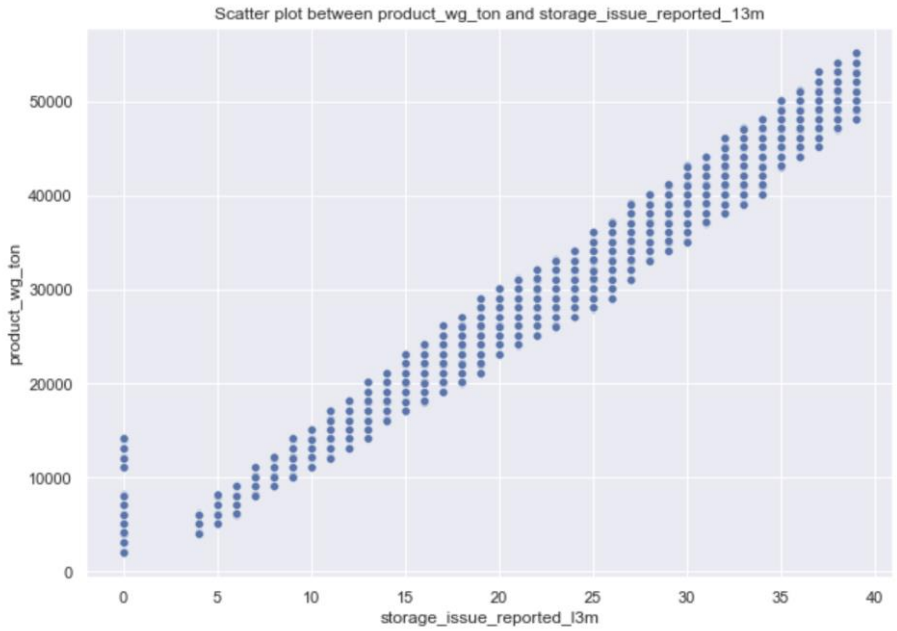
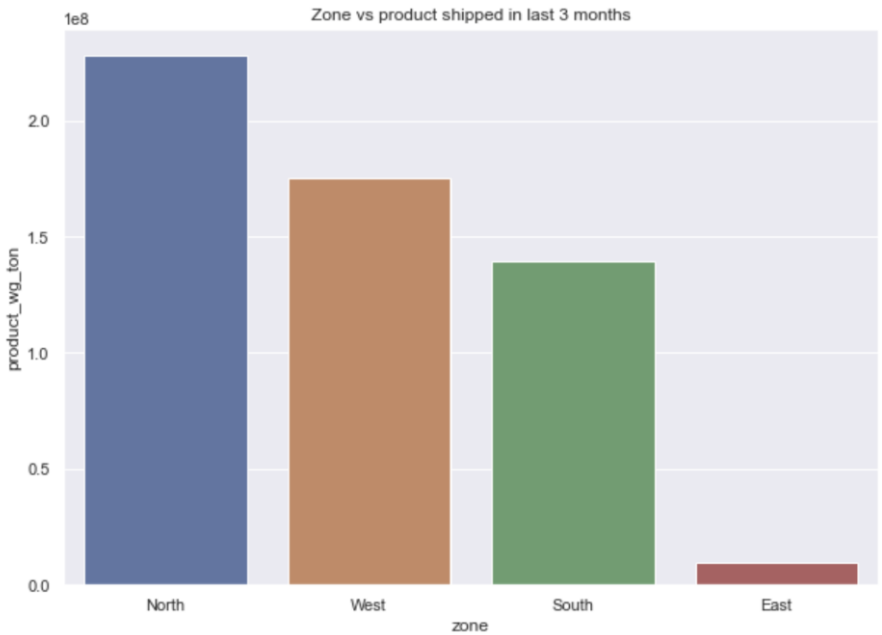


Figure 4 : Correlation matrix of the data

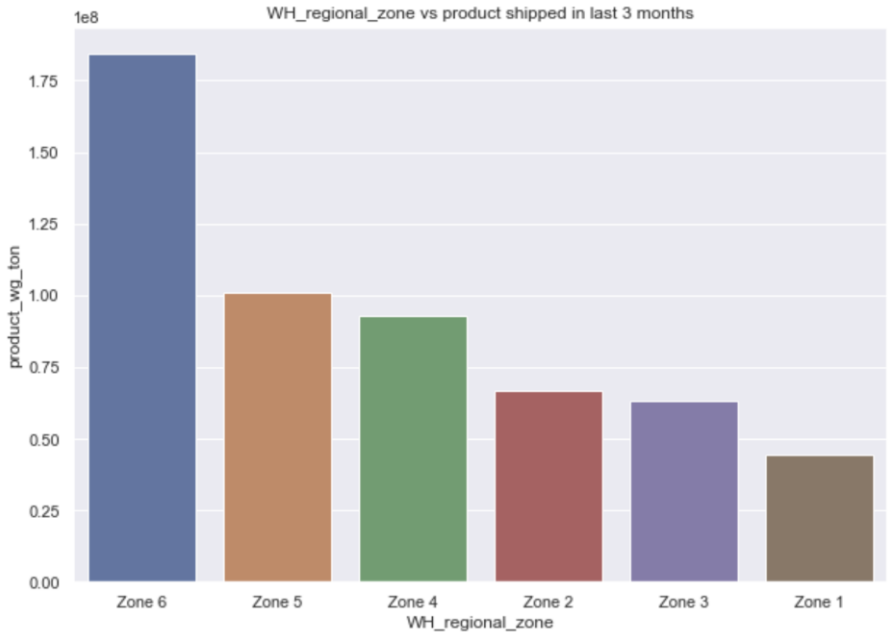
- The columns product_wg_ton and storage_issue_reported_13m have a strong positive correlation.
- The columns transport_issue_11y and product_wg_ton have a negative correlation.
- The columns transport_issue_11y and storage_issue_reported_13m have a negative correlation.
- The columns retail_shop_num and Competitor_in_mkl have a negative correlation.



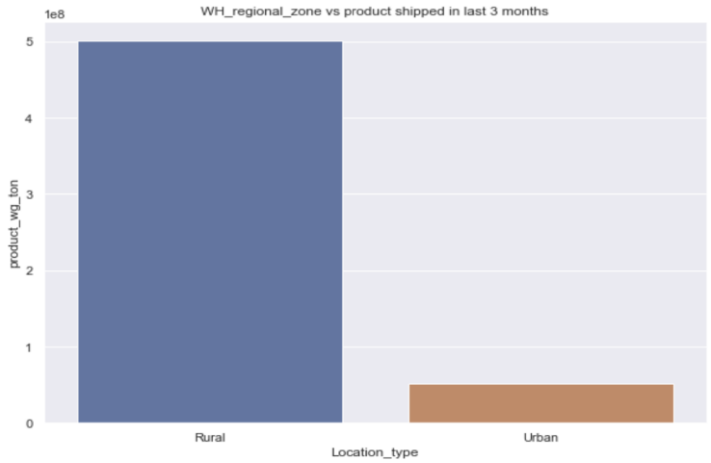
- The columns product_wg_ton and storage_issue_reported_13m have strong linear relationship.



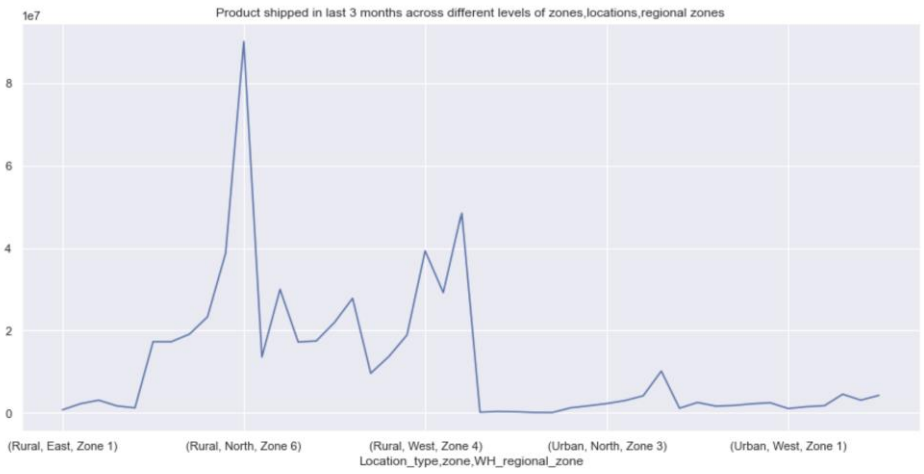
- The sum of product shipped across different zones are
  North – 228165823, West – 175111596, South – 139540901 and East – 9747503



- The sum of product shipped across different WH_regional_zone is Zone 6 – 184421651, Zone 5 – 101017613 , Zone 4 – 92596029, Zone 2 – 66580768 , Zone 3 – 63290230 , Zone 1 – 44659532



- The sum of product shipped across different Location_Type are Rural – 501482582, Urban – 5108321



- The sum of product shipped seems to be low in Urban areas when compared to rural areas.
- The Rural, North, Zone 6 have the highest sum of product shipped. Hence, advertisement campaigns can be done more in those areas to make the product more popular.
- The Urban, East and Zone 1 have the least sum of product shipped. Hence the management needs to identify the reasons for that take necessary measures such as providing offers and discounts.

3. Data Cleaning and Pre-processing

- The columns Ware_house_ID,WH_Manager_ID,wh_est_year are not necessary for further analysis. Hence, they can be dropped from the dataset.
- The columns worker_num and approved_wh_govt_certificate have null values
- The column worker_num is a numeric column. So, the null values are imputed with its median value of 28.
- The approved_wh_govt_certificate is an object column. So, the null values are imputed with its mode value of C.

	LL	Q1	Q2	Q3	UL	no_of_outliers	Outliers(%)	Outliers
num_refill_req_l3m	-4.0	2.0	4.0	6.0	12.0	0	0.000	[]
transport_issue_l1y	-1.5	0.0	0.0	1.0	2.5	2943	11.772	[4, 3, 3, 3, 4, 5, 3, 3, 3, 3, 3, 5, 4, 3, 3, ...
Competitor_in_mkt	-1.0	2.0	3.0	4.0	7.0	96	0.384	[8, 8, 10, 8, 8, 8, 9, 8, 10, 8, 9, 12, 8, 8, ...
retail_shop_num	2532.5	4313.0	4859.0	5500.0	7280.5	948	3.792	[7692, 7311, 8913, 8736, 8137, 7896, 8236, 805...
distributor_num	-11.5	29.0	42.0	56.0	96.5	0	0.000	[]
dist_from_hub	-54.5	109.0	164.0	218.0	381.5	0	0.000	[]
workers_num	10.5	24.0	28.0	33.0	46.5	607	2.428	[47.0, 48.0, 50.0, 62.0, 49.0, 49.0, 56.0, 53....
storage_issue_reported_l3m	-11.0	10.0	18.0	24.0	45.0	0	0.000	[]
wh_breakdown_l3m	-2.5	2.0	3.0	5.0	9.5	0	0.000	[]
govt_check_l3m	-11.5	11.0	21.0	26.0	48.5	0	0.000	[]
product_wg_ton	-12507.0	13059.0	22101.0	30103.0	55669.0	0	0.000	[]

- Outliers are present in transport_issue_l1y, Competitor_in_mkt, retailer_shop_num and workers_num.
- Hence, the outliers in the respective colums are replaced with their lower and upper limits.
- Lower limit = Q1 – 1.5 (IQR) , Upper = Q3 + 1.5 (IQR) , IQR = Q3 – Q1.
- No variable transformation is required for the data.
- No addition of new variables is required for the data.
- Before building the model, the data needs to be numeric. Hence, we need to perform encoding on the object columns.
- The columns 'Location_type','wh_owner_type','flood_impacted','flood_proof', 'electric_supply','temp_reg_mach' are one hot encoded.
- The columns 'zone','WH_regional_zone','approved_wh_govt_certificate' are label encoded.
- As the column 'WH_capacity_size' is ordinal in nature , Large is encoded with 3, Mid with 2 and Small with 1.
- The target variable which we have to predict using the model is '**product_wg_ton'.** The split the data into features(X) and target variable(y).

- Splitting the data into train and test data sets with a split ratio of (70:30) and random state of 1.
    - The shape of X_train is (17500,20).
    - The shape of y_train is (17500,1).
    - The shape of X_test is (7500,1).
    - The shape of y_test is (7500,1).
- As we are trying to build a regression model, if the ranges of the feature variables are varying it would affect the performance of the model, we perform normalization(min-max scaling) of the features in order to bring the range between 0 and 1.

4. Model Building

- **Model 1 -** Building a linear regression model using the train data set we obtain the following results, coefficients and results –

```
The Mean Absolute Error (MAE) of the model for Train set is 1294.5243727069453

The Root Mean Square Error (RMSE) of the model for Train set is 1771.8361446282172

The Mean Absolute Percentage Error (MAPE) of the model for Train set is 9.01%

The coefficient of determination R^2 of the prediction on Train set 0.9768484609882184

The Adjusted R^2 for Train set is 0.976821970297662

The coefficient for WH_capacity_size is -21.935480598392978
The coefficient for zone is -11.410633823054884
The coefficient for WH_regional_zone is -35.32693907291931
The coefficient for num_refill_req_l3m is -20.102488292155446
The coefficient for transport_issue_l1y is -1553.4577572276787
The coefficient for Competitor_in_mkt is -92.32759453133453
The coefficient for retail_shop_num is -110.95827876459073
The coefficient for distributor_num is 64.82054843959068
The coefficient for dist_from_hub is 56.823481604329764
The coefficient for workers_num is -19.919064040399206
The coefficient for storage_issue_reported_l3m is 48960.684779346724
The coefficient for approved_wh_govt_certificate is -420.2395218723305
The coefficient for wh_breakdown_l3m is -1469.5871649417659
The coefficient for govt_check_l3m is -7.288106794325037
The coefficient for Location_type_Urban is -108.94308630062469
The coefficient for wh_owner_type_Rented is 13.946614471534089
The coefficient for flood_impacted_1 is 21.192317025763828
The coefficient for flood_proof_1 is 54.54872401194445
The coefficient for electric_supply_1 is 11.130734990234977
The coefficient for temp_reg_mach_1 is 902.2517325864919

The intercept for the model is 1704.6257489040327
```

- **Model 2** – Linear regression with VIF treatment
    - In order to reduce the multicollinearity between the features, we apply variance inflation factor (VIF) treatment. The features that are having VIF value of more than 5 are discarded in the model building process.

    - After applying the VIF the following features are discarded –

        - retail_shop_num, workers_num, WH_capacity_size, Competitor_in_mkt, distributor_num, dist_from_hub, wh_breakdown_l3m.

    - The columns that are selected for building the model after VIF treatment are –

        - govt_check_l3m, zone, WH_regional_zone, storage_issue_reported_l3m, num_refill_req_l3m, approved_wh_govt_certificate, electric_supply_1, wh_owner_type_Rented, temp_reg_mach_1, transport_issue_l1y, flood_impacted_1, Location_type_Urban, flood_proof_1.

- Splitting the data into train and test data sets with a split ratio of (70:30) and random state of 1 after VIF treatment
  - The shape of X_train is (17500,13).
  - The shape of y_train is (17500,1).
  - The shape of X_test is (7500,13).
  - The shape of y_test is (7500,1).
- A regression model is trained on the above data and the results are as follows –

```
The Mean Absolute Error (MAE) of the model for Train set is 1300.4203511593867

The Root Mean Square Error (RMSE) of the model for Train set is 1812.4455354994773

The Mean Absolute Percentage Error (MAPE) of the model for Train set is 9.24%

The coefficient of determination R^2 of the prediction on Train set 0.9757750615581746

The Adjusted R^2 for Train set is 0.9757570514815566

The coefficient for zone is -26.002521879819017
The coefficient for WH_regional_zone is -46.60380336997041
The coefficient for num_refill_req_l3m is -15.820661226274137
The coefficient for transport_issue_l1y is -1650.1457547265816
The coefficient for storage_issue_reported_l3m is 48311.93909754236
The coefficient for approved_wh_govt_certificate is -313.1306304571404
The coefficient for govt_check_l3m is 0.2672098694038823
The coefficient for Location_type_Urban is -134.2534615046218
The coefficient for wh_owner_type_Rented is 5.247890355981313
The coefficient for flood_impacted_1 is 27.57655485504869
The coefficient for flood_proof_1 is 74.87764481750972
The coefficient for electric_supply_1 is 16.821650473898973
The coefficient for temp_reg_mach_1 is 886.5397792745314

The intercept for the model is 1091.4380503637549
```

- **Model 3 – Random forest Regressor(Ensemble modeling)**

  - As the problem statement deals with the regression problem, we can use **RandomForestRegressor** as part of ensemble modelling.
  - Once the RandomForestRegressor is used to train the data the following results are obtained on the train data.

```
The Mean Absolute Error (MAE) of the model for Train set is 265.21680171428574

The Root Mean Square Error (RMSE) of the model for Train set is 362.51392596715505

The Mean Absolute Percentage Error (MAPE) of the model for Train set is 1.71%

The coefficient of determination R^2 of the prediction on Train set 0.999030869767335

The Adjusted R^2 for Train set is 0.9990297608592366
```

- **Model 4 - Random forest Regressor using GridSearchCV**

```
The best parameters are {'n_estimators': 20}

The best estimator is : RandomForestRegressor(n_estimators=20)

Best estimator's score on training data is 99.89%

The Mean Absolute Error (MAE) of the model for Train set is 275.93964

The Root Mean Square Error (RMSE) of the model for Train set is 390.77939478212727

The Mean Absolute Percentage Error (MAPE) of the model for Train set is 1.79%

The coefficient of determination R^2 of the prediction on Train set 0.9988738504581226

The Adjusted R^2 for Train set is 0.9988725618837856
```

- **Model 5 - Random forest Regressor using GridSearchCV**

```
The best parameters are {'criterion': 'squared_error', 'n_estimators': 20}
The best estimator is : RandomForestRegressor(n_estimators=20)
Best estimator's score on training data is 99.89%
The Mean Absolute Error (MAE) of the model for Train set is 276.38111142857144
The Root Mean Square Error (RMSE) of the model for Train set is 391.0846052452808
The Mean Absolute Percentage Error (MAPE) of the model for Train set is 1.78%
The coefficient of determination R^2 of the prediction on Train set 0.998872090657822
The Adjusted R^2 for Train set is 0.9988708000698683
```

- **Model 6 - Random forest Regressor using GridSearchCV**

```
The best parameters are {'criterion': 'squared_error', 'max_depth': 7, 'n_estimators': 20}
The best estimator is : RandomForestRegressor(max_depth=7, n_estimators=20)
Best estimator's score on training data is 99.23%
The Mean Absolute Error (MAE) of the model for Train set is 751.2900465225732
The Root Mean Square Error (RMSE) of the model for Train set is 1020.1866484600336
The Mean Absolute Percentage Error (MAPE) of the model for Train set is 4.75%
The coefficient of determination R^2 of the prediction on Train set 0.9923247589562991
The Adjusted R^2 for Train set is 0.9923159767135579
```

- **Model 7 - Random forest Regressor using GridSearchCV**

```
The best parameters are {'criterion': 'squared_error', 'max_depth': 7, 'min_samples_leaf': 4, 'n_estimators': 50}
The best estimator is : RandomForestRegressor(max_depth=7, min_samples_leaf=4, n_estimators=50)
Best estimator's score on training data is 99.23%
The Mean Absolute Error (MAE) of the model for Train set is 751.0659613926209
The Root Mean Square Error (RMSE) of the model for Train set is 1020.2698368667437
The Mean Absolute Percentage Error (MAPE) of the model for Train set is 4.74%
The coefficient of determination R^2 of the prediction on Train set 0.9923235071910335
The Adjusted R^2 for Train set is 0.9923147235159846
```

5. Model Validation

- All the above built models are validated on test dataset and the results are as follows-
- Model 1 –

```
The Mean Absolute Error (MAE) of the model for Test set is 1275.0808234962778
The Root Mean Square Error (RMSE) of the model for Test set is 1711.5804472171883
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 8.84%
The coefficient of determination R^2 of the prediction on Test set 0.9779245836515131
The Adjusted R^2 for Test set is 0.9778655505819892
```

- Model 2 –

```
The Mean Absolute Error (MAE) of the model for Test set is 1276.104174643809
The Root Mean Square Error (RMSE) of the model for Test set is 1744.02384993421
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 9.08%
The coefficient of determination R^2 of the prediction on Test set 0.9770797627054585
The Adjusted R^2 for Test set is 0.9770399599957565
```

- Model 3 –

```
The Mean Absolute Error (MAE) of the model for Test set is 701.9361093333332
The Root Mean Square Error (RMSE) of the model for Test set is 940.0584865458106
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 4.55%
The coefficient of determination R^2 of the prediction on Test set 0.9933407699385448
The Adjusted R^2 for Test set is 0.9933229621298499
```

- Model 4 –

```
The Mean Absolute Error (MAE) of the model for Test set is 713.2514733333334
The Root Mean Square Error (RMSE) of the model for Test set is 959.907599151953
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 4.62%
The coefficient of determination R^2 of the prediction on Test set 0.9930565848969612
The Adjusted R^2 for Test set is 0.9930380171336157
```

- Model 5 –

```
The Mean Absolute Error (MAE) of the model for Test set is 716.34844
The Root Mean Square Error (RMSE) of the model for Test set is 959.5212752277391
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 4.64%
The coefficient of determination R^2 of the prediction on Test set 0.9930621726589338
The Adjusted R^2 for Test set is 0.9930436198381261
```

- Model 6 –

```
The Mean Absolute Error (MAE) of the model for Test set is 745.279583116128
The Root Mean Square Error (RMSE) of the model for Test set is 989.7503843985943
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 4.65%
The coefficient of determination R^2 of the prediction on Test set 0.9926181429917302
The Adjusted R^2 for Test set is 0.9925984027670791
```

- Model 7 –

```
The Mean Absolute Error (MAE) of the model for Test set is 744.6886785939761
The Root Mean Square Error (RMSE) of the model for Test set is 988.9120095677337
The Mean Absolute Percentage Error (MAPE) of the model for Test set is 4.63%
The coefficient of determination R^2 of the prediction on Test set 0.9926306434001184
The Adjusted R^2 for Test set is 0.9926109366034882
```

	MAE	RMSE	MAPE(%)	R2(%)	Adjusted R2(%)
Model 1	1294.52	1771.83	9.01	97.68	97.68
Model 2	1300.42	1812.44	9.24	97.57	97.57
Model 3	265.17	363.00	1.71	99.90	99.90
Model 4	275.93	390.77	1.79	99.88	99.88
Model 5	276.38	391.08	1.78	99.88	99.88
Model 6	751.29	1020.18	4.75	99.23	99.23
Model 7	751.06	1020.26	4.74	99.23	99.23

Table 3 : Validation metrics for train data set

	MAE	RMSE	MAPE(%)	R2(%)	Adjusted R2(%)
Model 1	1275.08	1711.58	8.84	97.79	97.78
Model 2	1276.10	1744.02	9.08	97.70	97.70
Model 3	701.93	940.05	4.55	99.33	99.33
Model 4	713.25	959.90	4.62	99.30	99.30
Model 5	716.34	959.59	4.64	99.30	99.30
Model 6	745.27	989.75	4.65	99.26	99.25
Model 7	744.68	988.91	4.63	99.26	99.26

Table 4 : Validation metrics for test data set

- From the above model comparisons, we can identify that the model 7 performs better than the other models as it performs almost same on train and test sets.
- Hence, model 7 is chosen for the prediction.

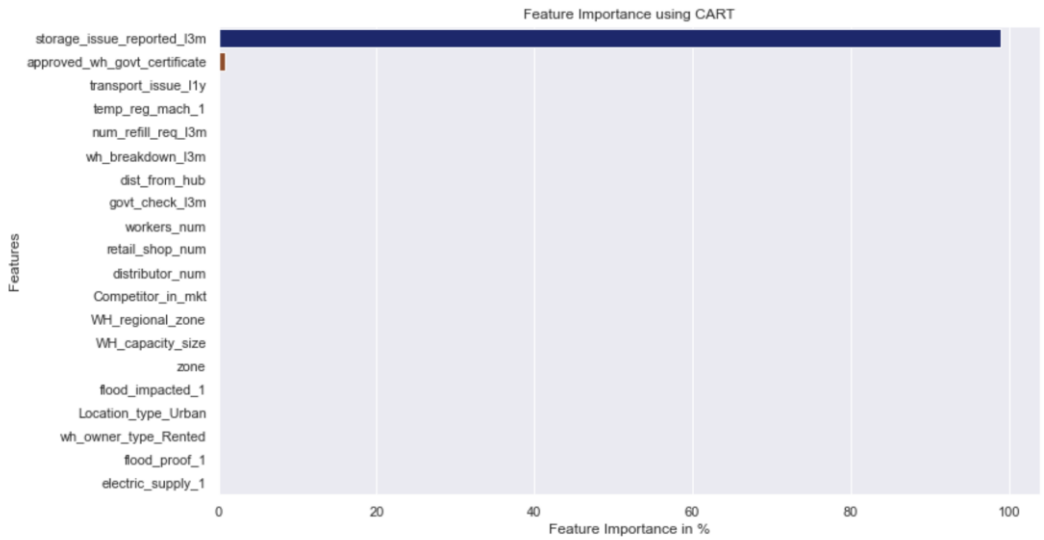


Figure 5 : Feature Importance of the best model

6. Final Interpretation/Recommendation

Interpretations

- The columns WH_capacity_size, zone, WH_regional_zone, num_refill_req_l3m, transport_sue_l1y, Competitor_in_mkt, retail_shop_num, workers_num, approved_wh_govt_certificate, wh_breakdown_l3m, govt_check_l3m, Location_type_Urban have a negative linear relationship with the target variable produc_wg_ton.

- The columns storage_issue_reported_l3m, wh_owner_type_Rented, flood_impacted_1, flood_proof_1, electric_supply_1, temp_reg_mach_1, dtributor_num, dt_from_hub have a positive linear relationship with the target variable product_wg_ton.

- The model has a good R^2 and adjusted R^2 score close to 1, which would indicate that the model can be used to predict the optimal product_wg_ton to be shipped.

- Of the columns that have negative linear relation with the target variable the columns transport_sue_l1y, wh_breakdown_l3m, approved_wh_govt_certificate have higher magnitude which would indicate that if there are any transport issues or warehouse break downs the shipment has a problem. And also, the approved govt certificate matters for the shipment.

- Of the columns that have positive linear relation with the target variable the columns storage_issue_reported_l3m, temp_reg_mach_1 have higher magnitude which would indicate that the warehouses that reported storage issues more the shipment was smooth.

- Once the VIF treatment is performed some features are discarded and the following columns govt_check_l3m, zone, WH_regional_zone, storage_issue_reported_l3m, num_refill_req_l3m, approved_wh_govt_certificate, electric_supply_1, wh_owner_type_Rented, temp_reg_mach_1, transport_issue_l1y, flood_impacted_1, Location_type_Urban, flood_proof_1 are considered for the optimal value to be shipped.

- The sum of product shipped seems to be low in Urban areas when compared to rural areas.
- The Rural, North, Zone 6 have the highest sum of product shipped.
- The Urban, East and Zone 1 have the least sum of product shipped.
- The model has identified that the columns storage_issue_reported_l3m, approved_wh_govt_certificate are the most important features.

Recommendations

- As the regression models have identified if there are any transport issues or warehouse breakdowns the shipment has some effect in those areas.
- Hence, if suppose the company plans to ship its shipment it should check the areas where there are transport issues or ware house breakdowns and avoid sending the shipment to those areas.
- And also, the company needs to take appropriate measures to avoid any transport issues and warehouse break downs in those areas.
- The Random Forest model has identified that the areas where the storage issues are reported has some relation with the shipment, the company can identify such areas and take necessary measures such as contact the management at the warehouses and identify the reason for the issues, to avoid such storage issues.

- The Random Forest model has also identified that the approved govt certificate has some relation with the shipment, the company can identify such areas and take the necessary measures like upgrading the certificate of ware houses to next higher level so that it doesn't affect the shipment.
- As the models have identified some areas where the shipment is affected and the reasons for it, the company can make advertisement campaigns in the areas where there are no storage issues, transport issue or warehouse break downs so that the company can gain profits.
- The Urban, East and Zone 1 have the least sum of product shipped. Hence the management needs to identify the reasons for that take necessary measures such as providing offers such as "buy 2 get 1 free" and discounts such as 50% or 60%.

THE END