Final PPT - Capstone

SUPPLY CHAIN PROJECT

SUBMITTED BY

SUNDAR RAM S

PGP DSBA –ONLINE DEC_C 2021

AGENDA

✓ Problem Definition & Objective

✓ Approach

✓ Insights & Recommendations

PROBLEM DEFINITION ?

- ✓Miss-Match in the demand and supply of Noodles
- ✓High Demand – Low Supply & Low Demand – High Supply
- ✓Inventory cost loss to the company

OBJECTIVE

Optimization of Supply Quantity in all ware houses

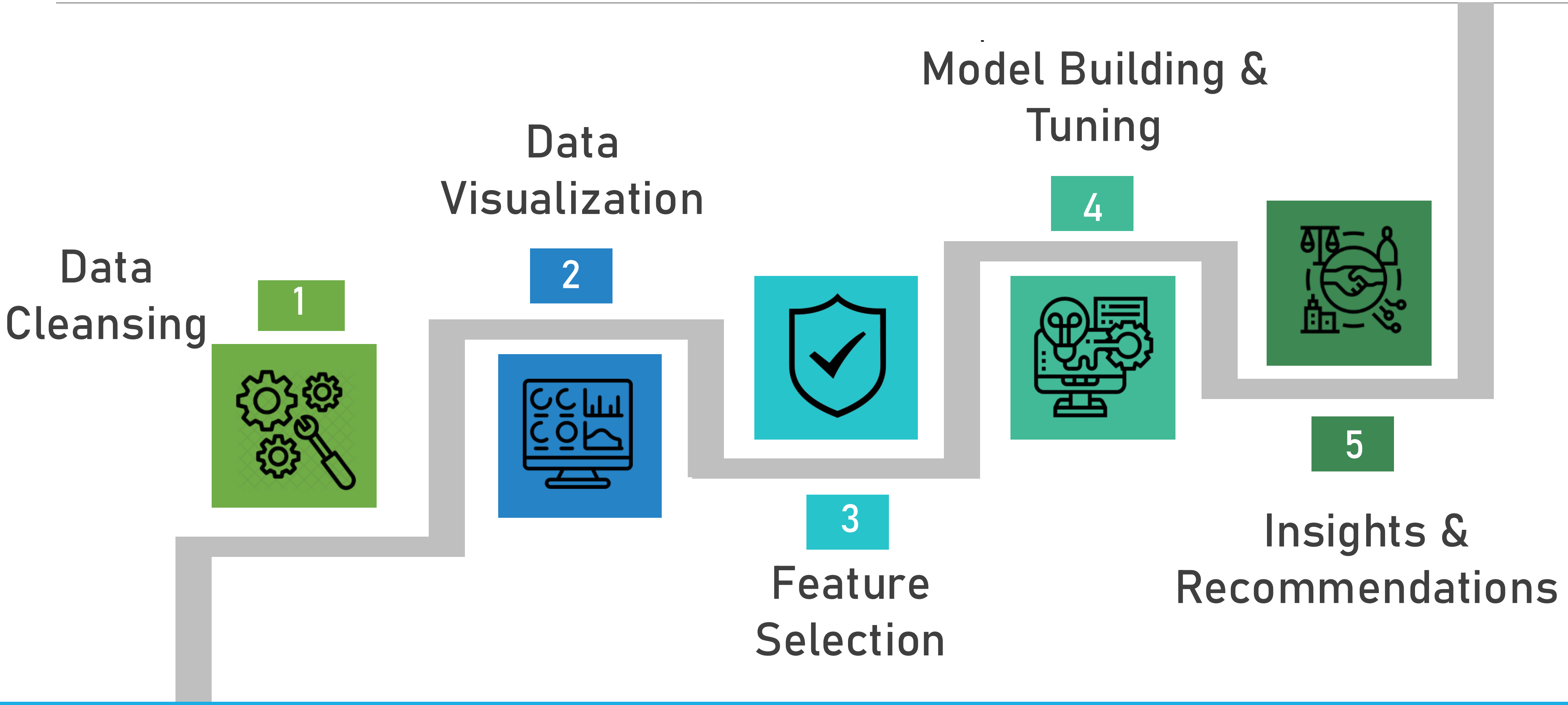
Determining optimum weight to be shipped

Analyze demand patterns

Boost sales/ bottom line through targeted campaigning

"80% of the cost is accommodated by only 20% of the products"

APPROACH


Data Cleansing
1
Data Visualization
2
3
Feature Selection
Model Building & Tuning
4
5
Insights & Recommendations

DATA OVERVIEW

- ✓ Total Records – 25K
- ✓ Total Variables – 24
- ✓ Total Cells – 6L
- ✓ 23 Independent Variables
- ✓ 1 dependent Target Variable

14 int64, 8 object type &
2 float64 variables

S.No	Variable Name	Data Type
1	Ware_house_ID	Object
2	WH_Manager_ID	Object
3	Location_type	Object
4	WH_capacity_size	Object
5	zone	Object
6	WH_regional_zone	Object
7	num_refill_req_l3m	int64
8	transport_issue_l1y	int64
9	Competitor_in_mkt	int64
10	retail_shop_num	int64
11	wh_owner_type	Object
12	distributor_num	int64
13	flood_impacted	int64
14	flood_proof	int64
15	electric_supply	int64
16	dist_from_hub	int64
17	workers_num	float64
18	wh_est_year	float64
19	storage_issue_reported_l3m	int64
20	temp_reg_mach	int64
21	approved_wh_govt_certificate	Object
22	wh_breakdown_l3m	int64
23	govt_check_l3m	int64
24	product_wg_ton	int64

DATA CLEANSING

Null Values

- Count: ~ 13.8 K (2% of the dataset)
- Most ML Algorithms don't work & Leads to Biased Models giving incorrect results
- Lack of Precision in Statistical Analysis
- Treated Using Forward Fill Technique

990 - workers_num
11881 - wh_est_year
908 - approved_wh_govt_certificate

Outliers

- Values abnormally away from the other values – Count: ~ 1.6 K (0.2% of the dataset)
- Affects arithmetic mean of the continuous variables & skews the value to one side
- Visualized using Box Plot
- Treated by imputing the max and min values

Variables	Minimum	Maximum
workers_num	10.5	46.5
retail_shop_num	2532.5	7280.5

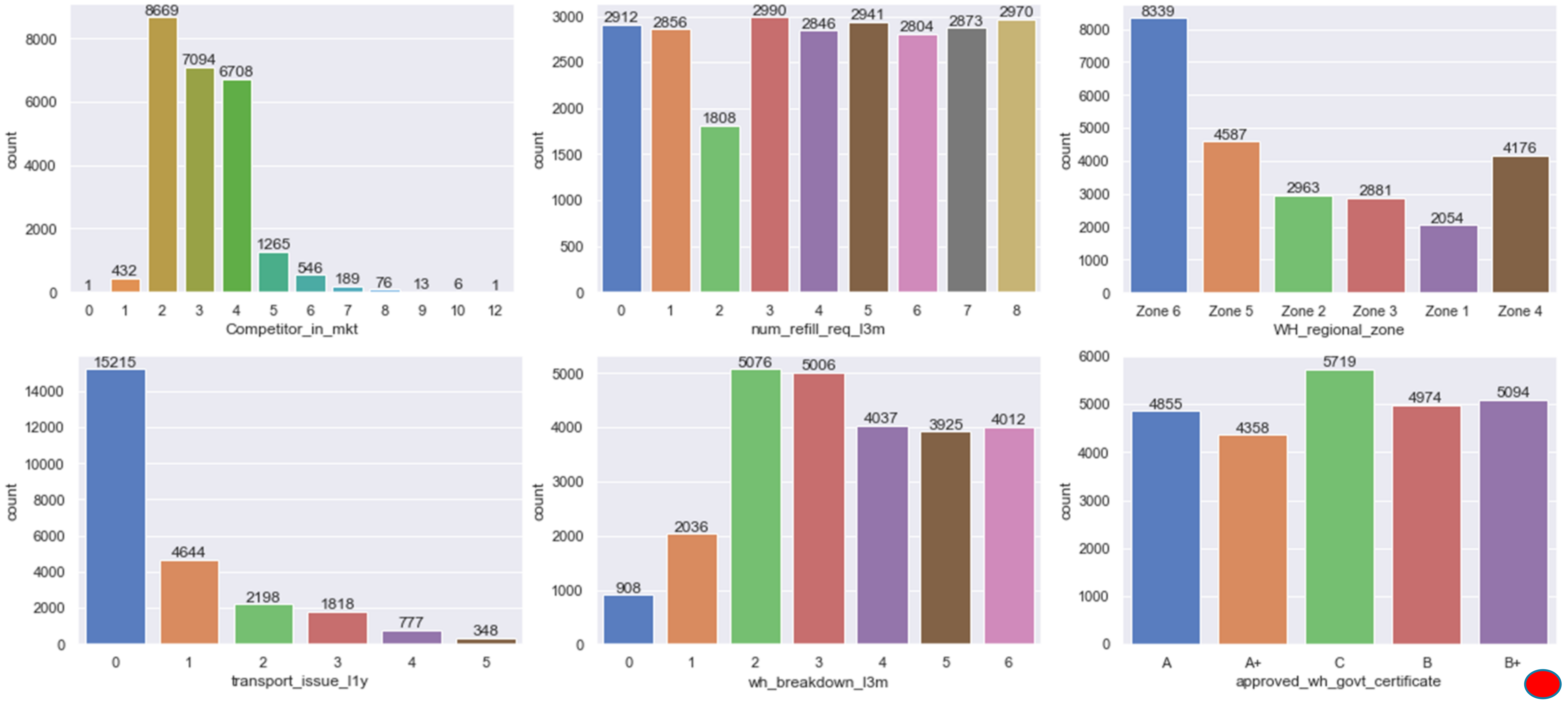
948 - retail_shop_num
631 - workers_num
Maximum = [Q3 + 1.5(IQR)]
Minimum = [Q1 - 1.5(IQR)]

Feature Engineering

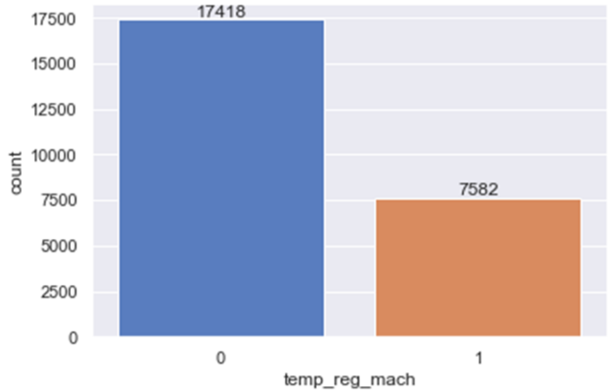
- Addition of New variables – Age of the Ware house – Added based on the wh_est_year
- Variable Transformation – Binning of Age & Weight variables
- Feature Engineering helps better Analysis & also sometimes better model building

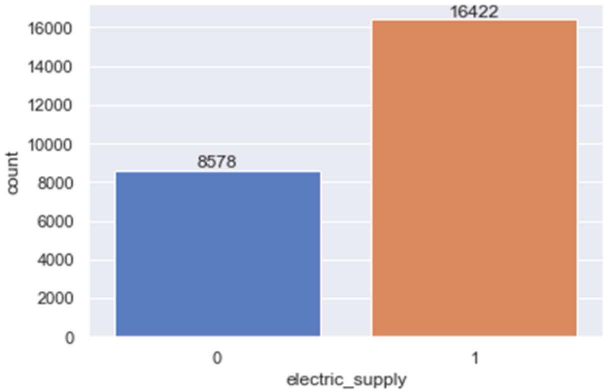
`Ware_house_ID & Manager_ID are dropped as they are mere identification numbers and don't add any value to model building`

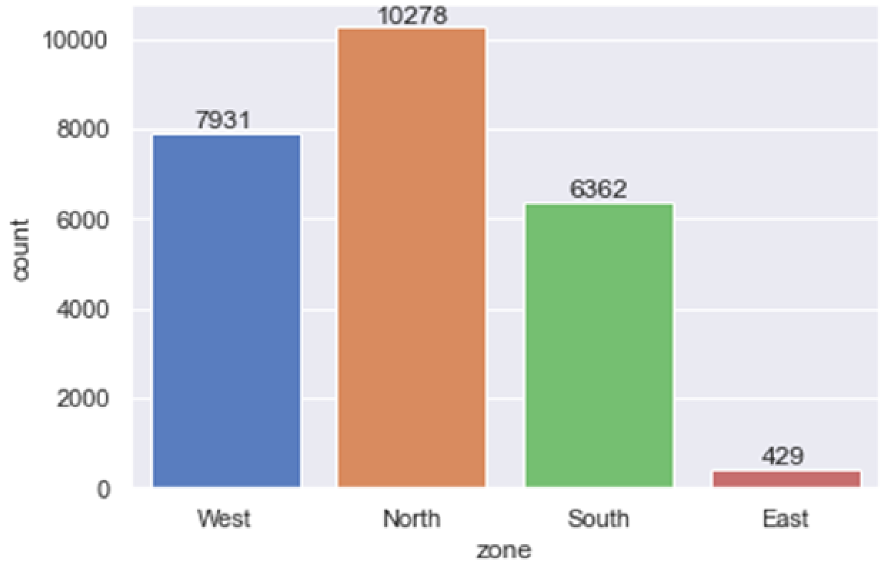
DATA VISUALIZATION

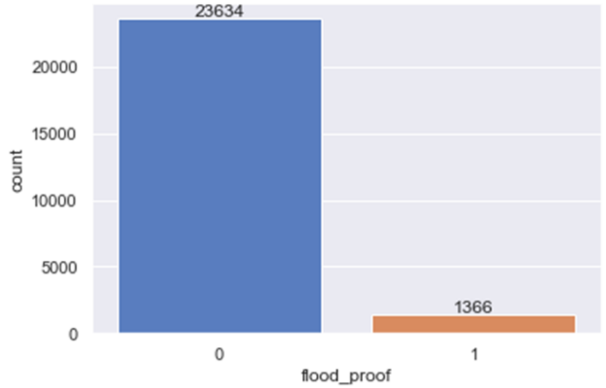


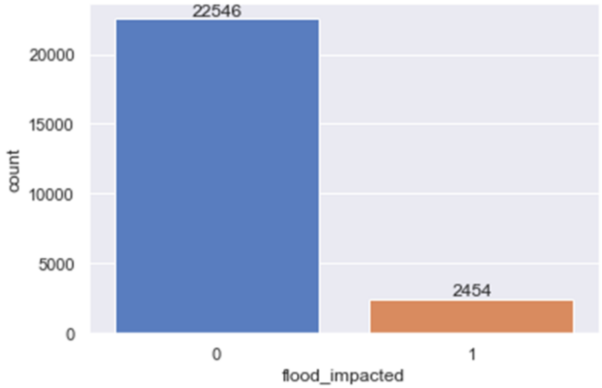
DATA VISUALIZATION – CONTD…

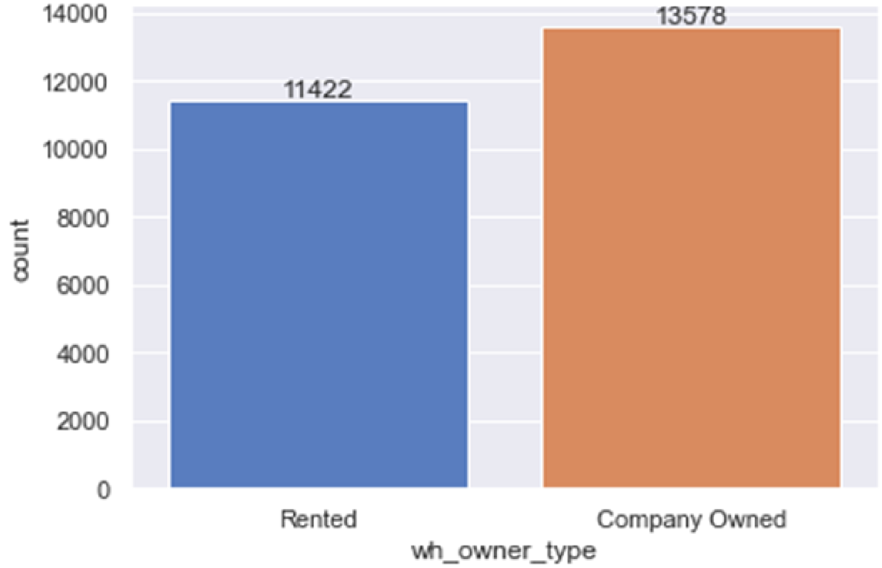


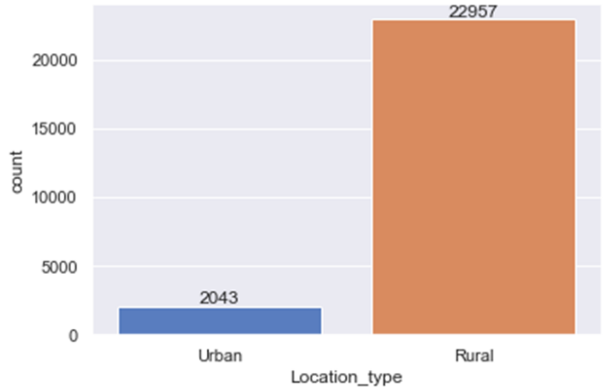


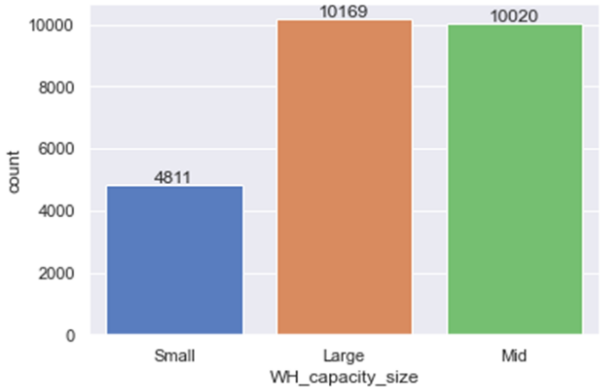




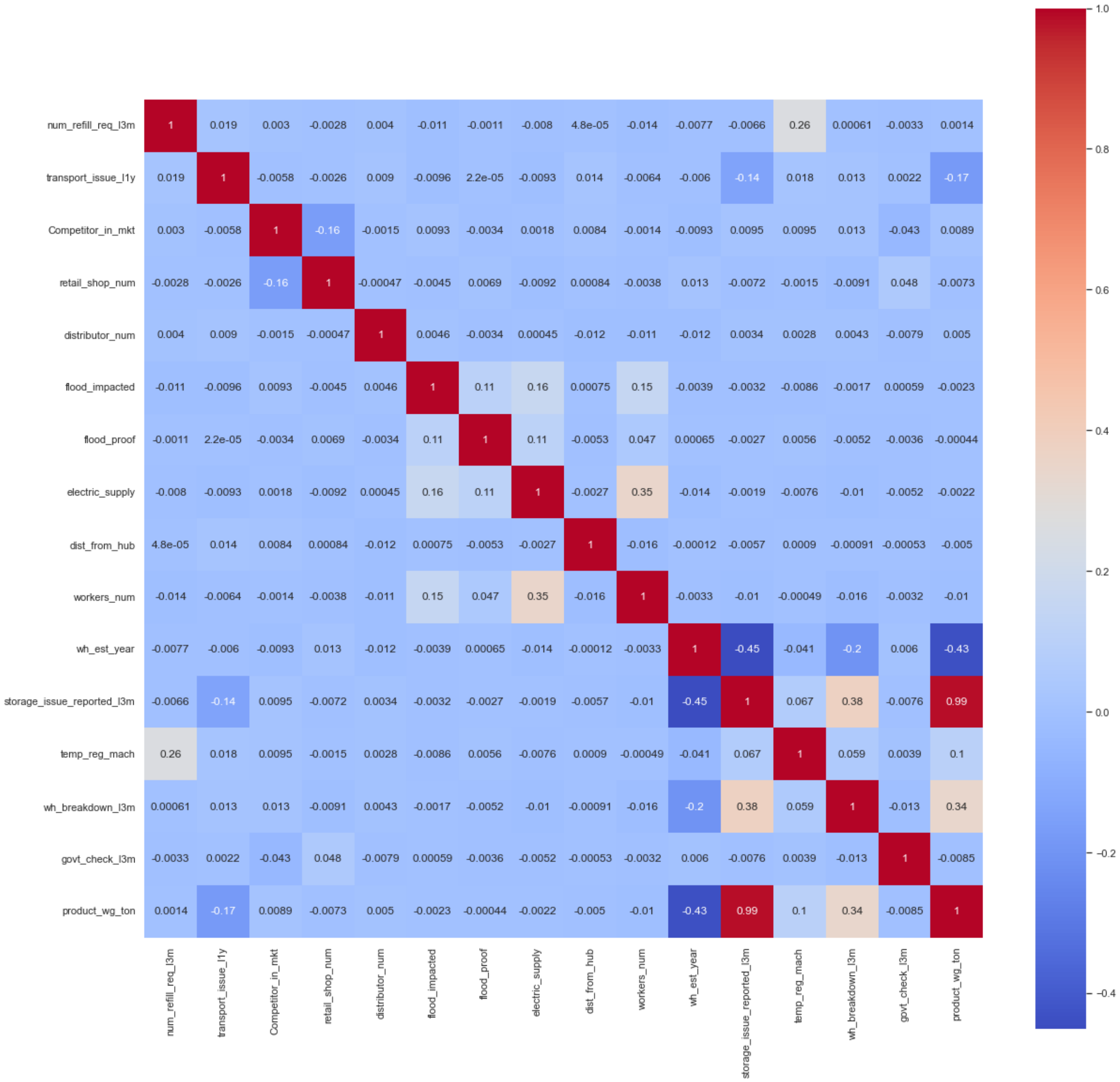






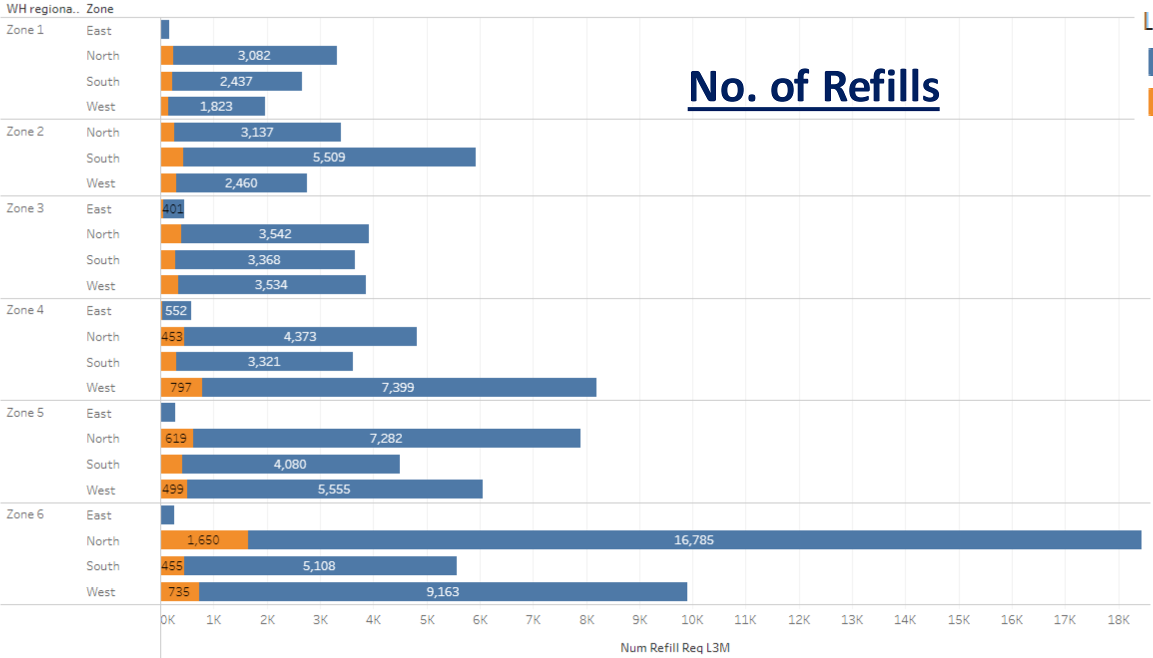


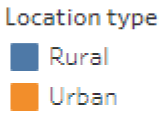
DATA VISUALIZATION – CONTD…



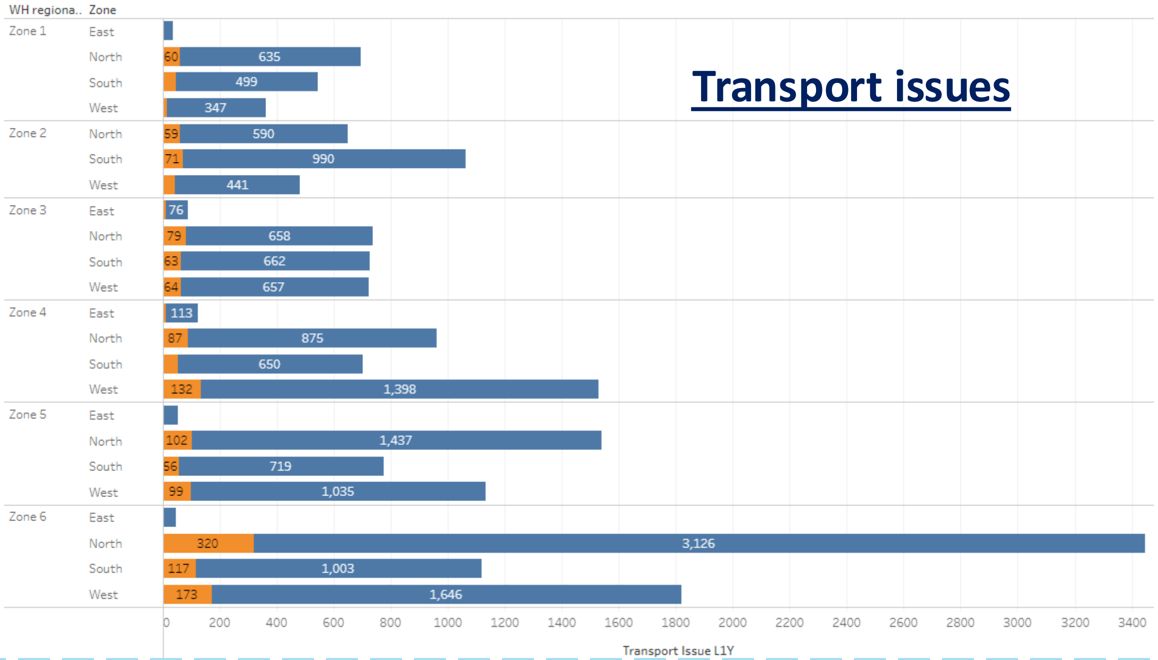
- ✓ Presence of Multi collinearity
- ✓ Indicates presence of correlation amongst independent variables
- ✓ Undermines the statistical significance of an independent variable
- ✓ Requires Suitable treatment

No. of Refills

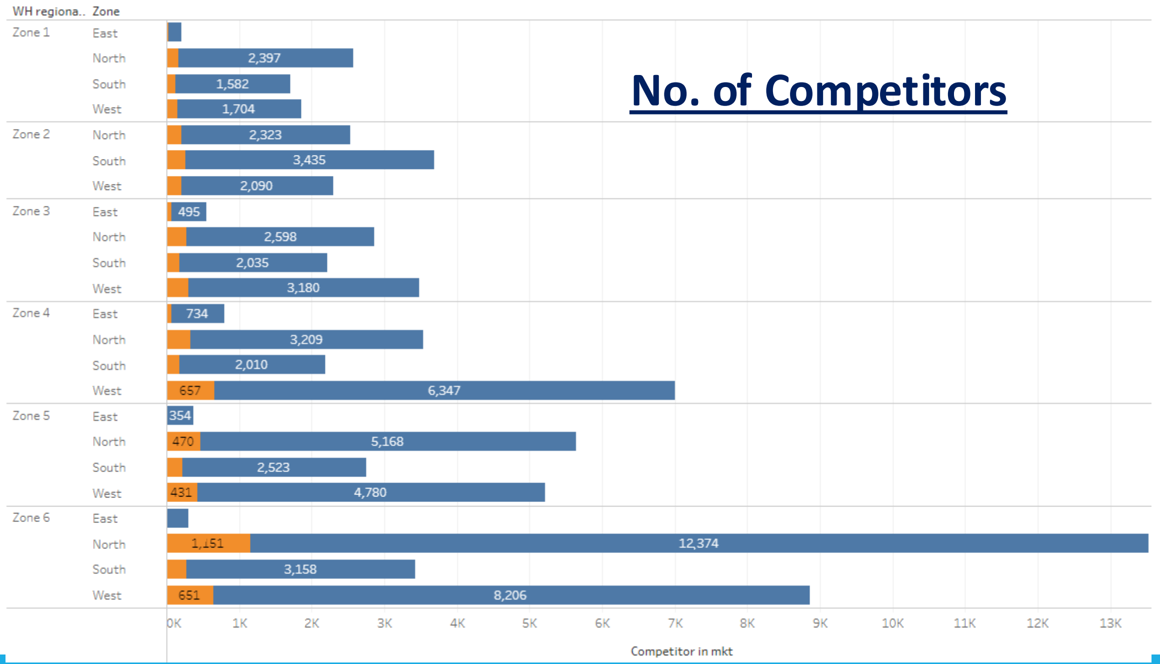




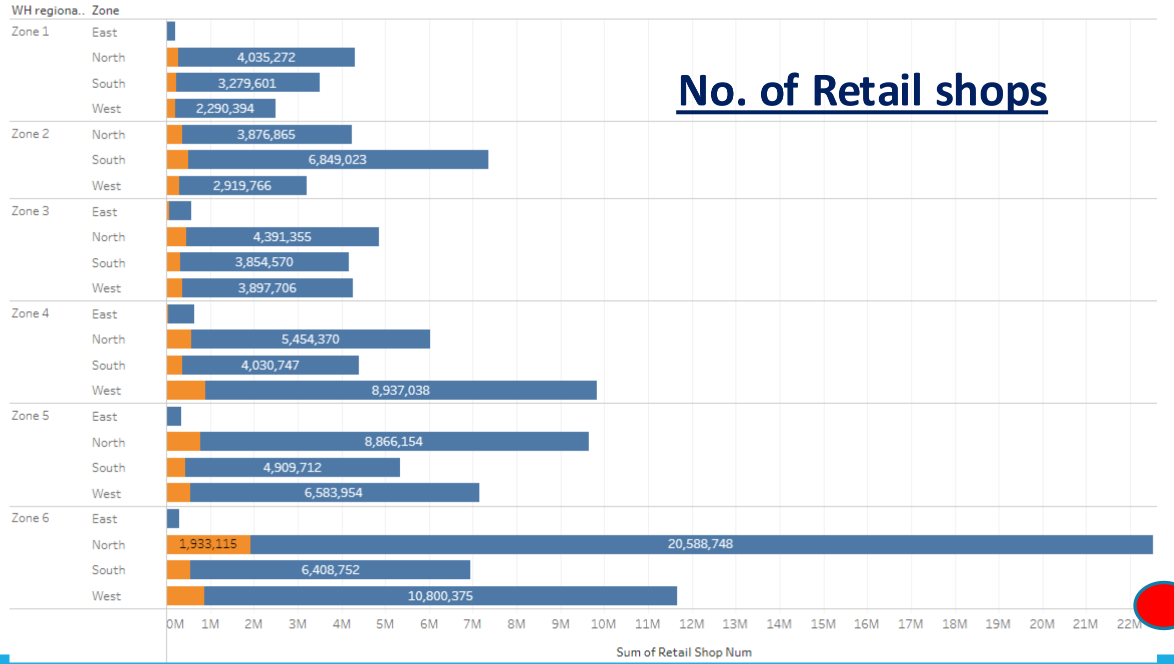
Transport issues

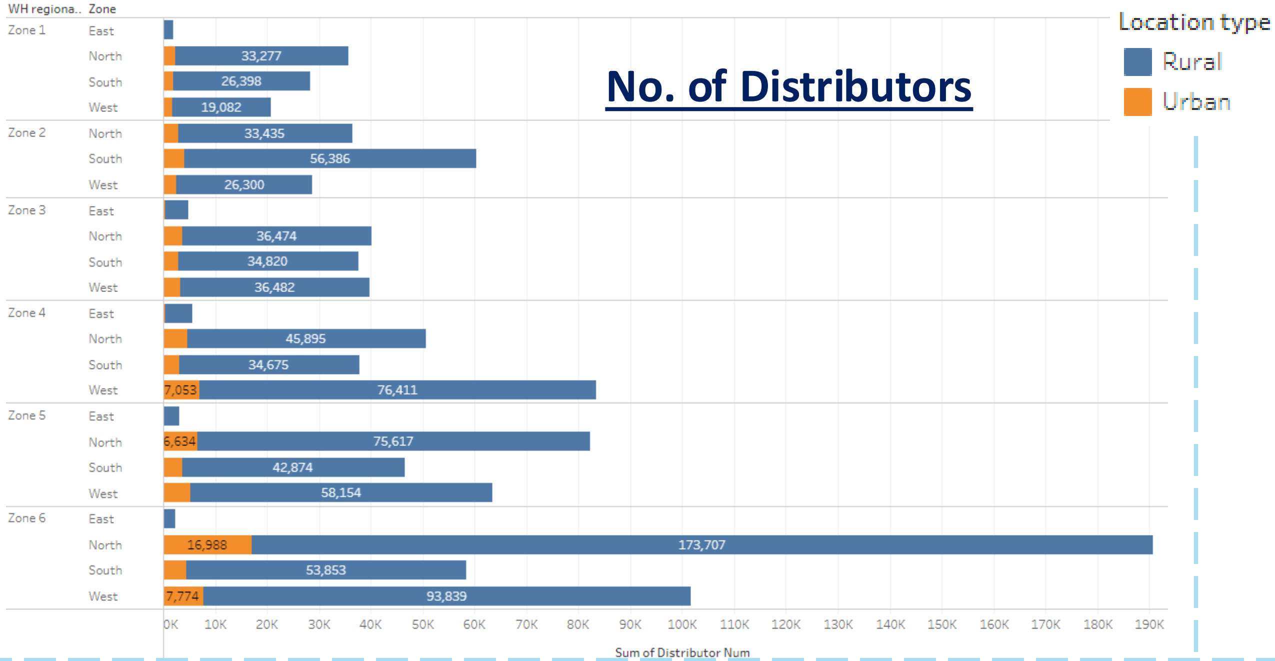


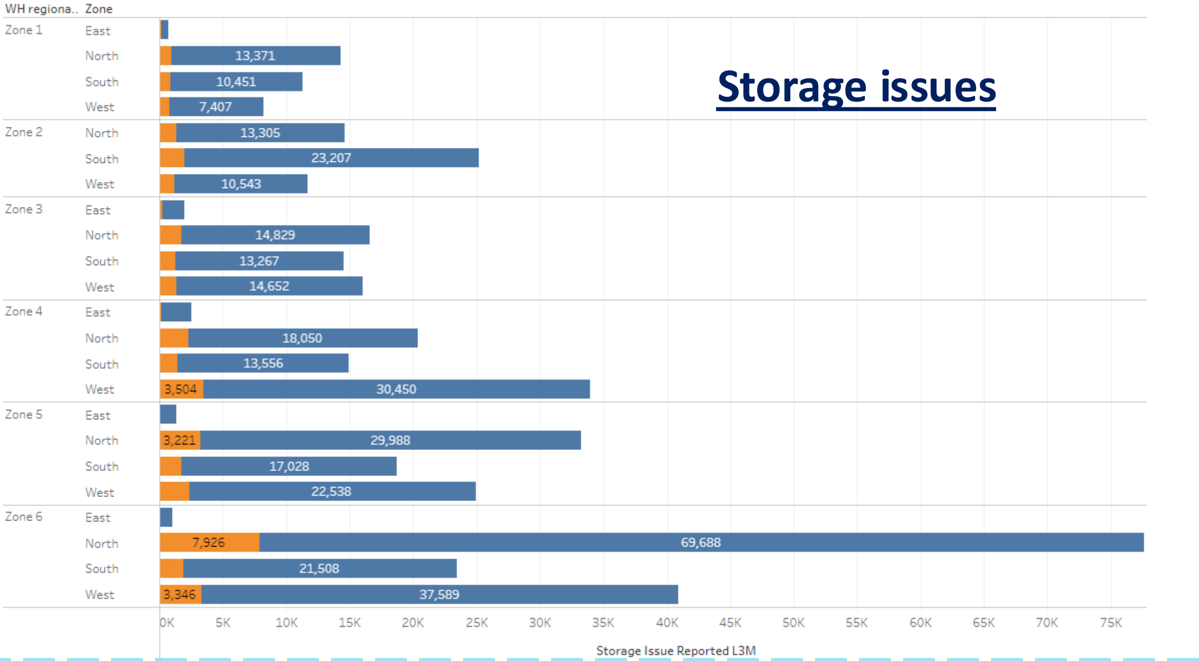
No. of Competitors

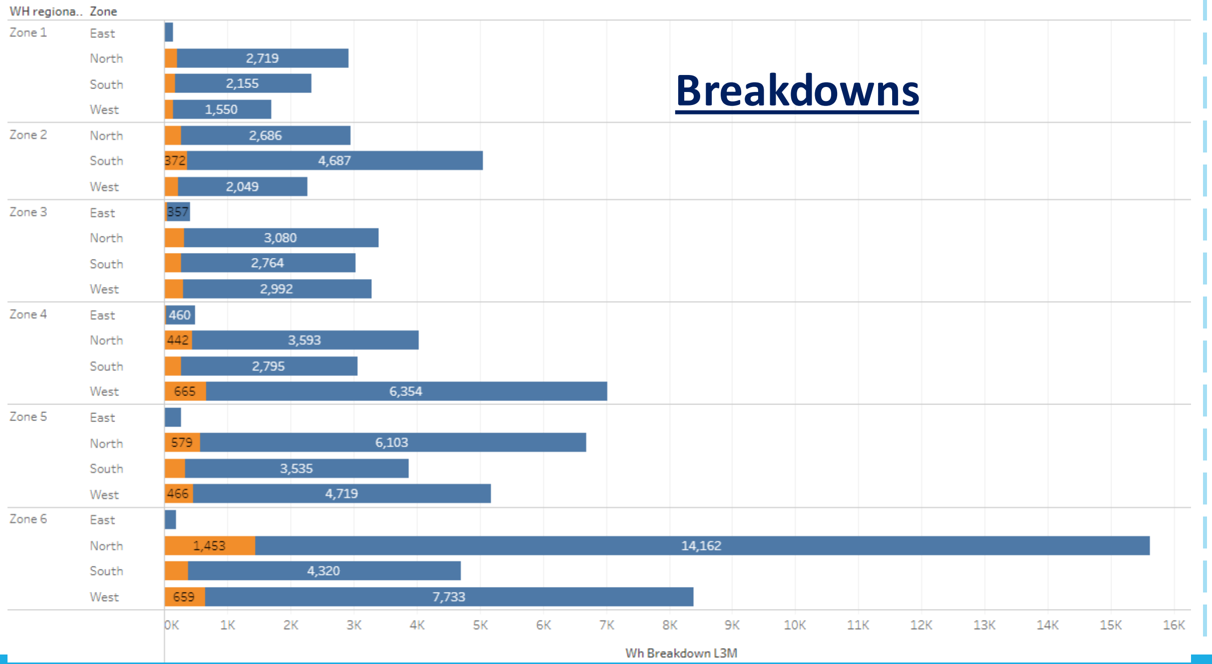


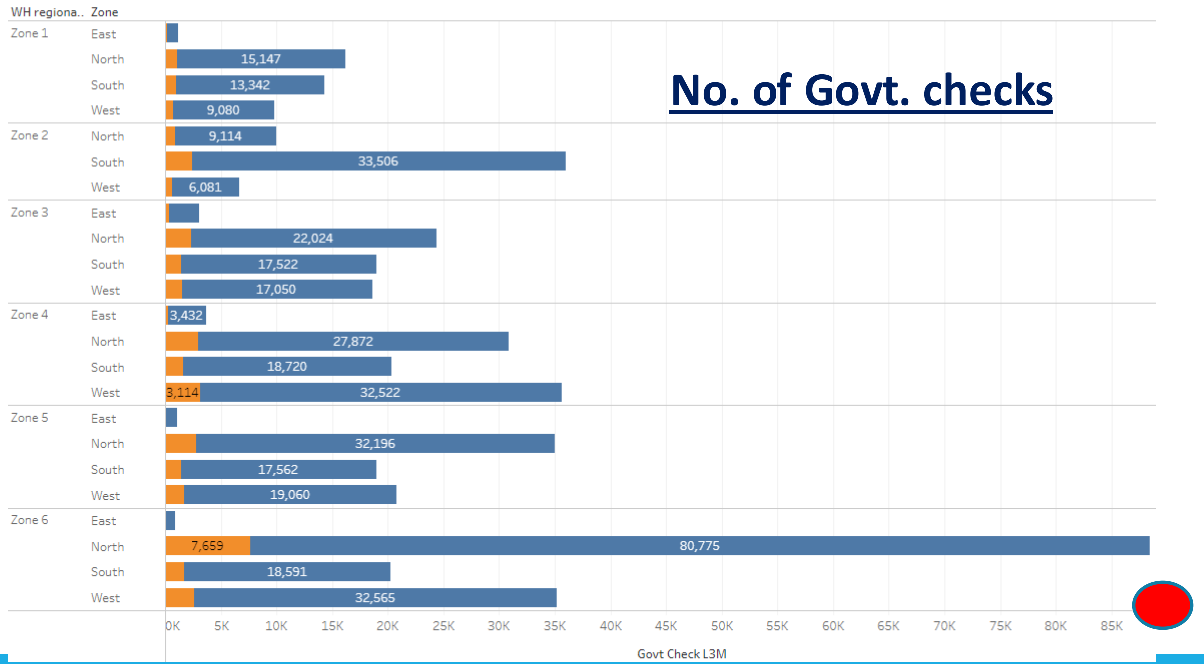
No. of Retail shops



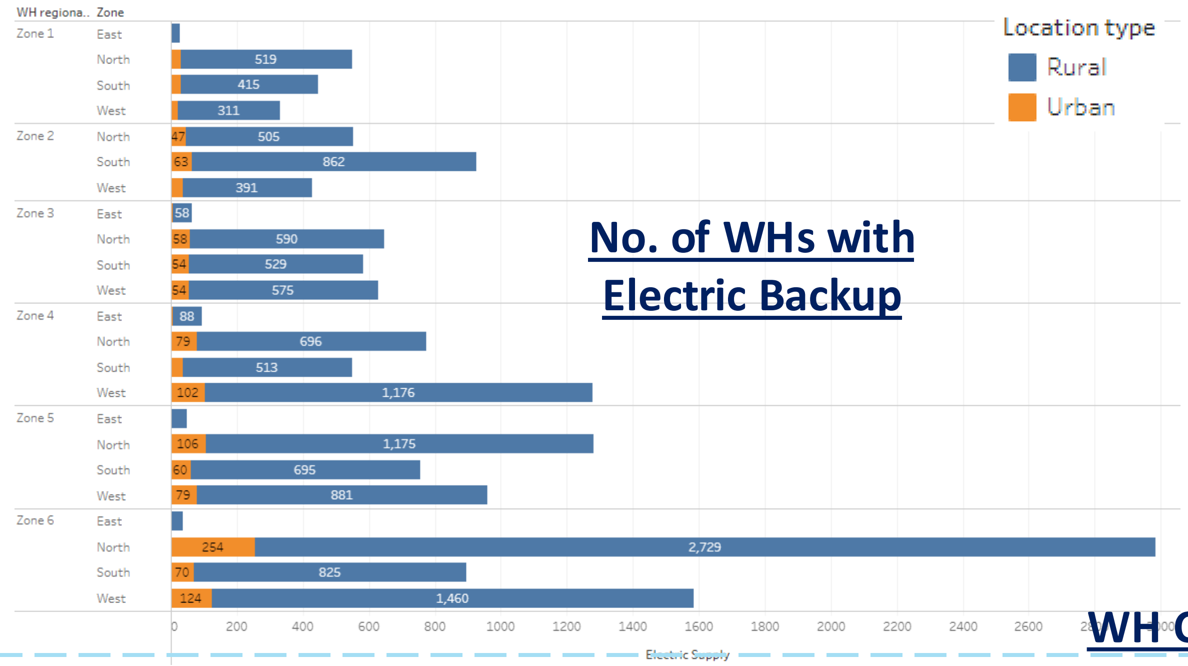

No. of Distributors
Location type
Rural
Urban
WH regiona.. Zone
Zone 1 East
North 33,277
South 26,398
West 19,082
Zone 2 North 33,435
South 56,386
West 26,300
Zone 3 East
North 36,474
South 34,820
West 36,482
Zone 4 East
North 45,895
South 34,675
West 7,053 76,411
Zone 5 East
North 6,634 75,617
South 42,874
West 58,154
Zone 6 East
North 16,988 173,707
South 53,853
West 7,774 93,839
0K 10K 20K 30K 40K 50K 60K 70K 80K 90K 100K 110K 120K 130K 140K 150K 160K 170K 180K 190K
Sum of Distributor Num

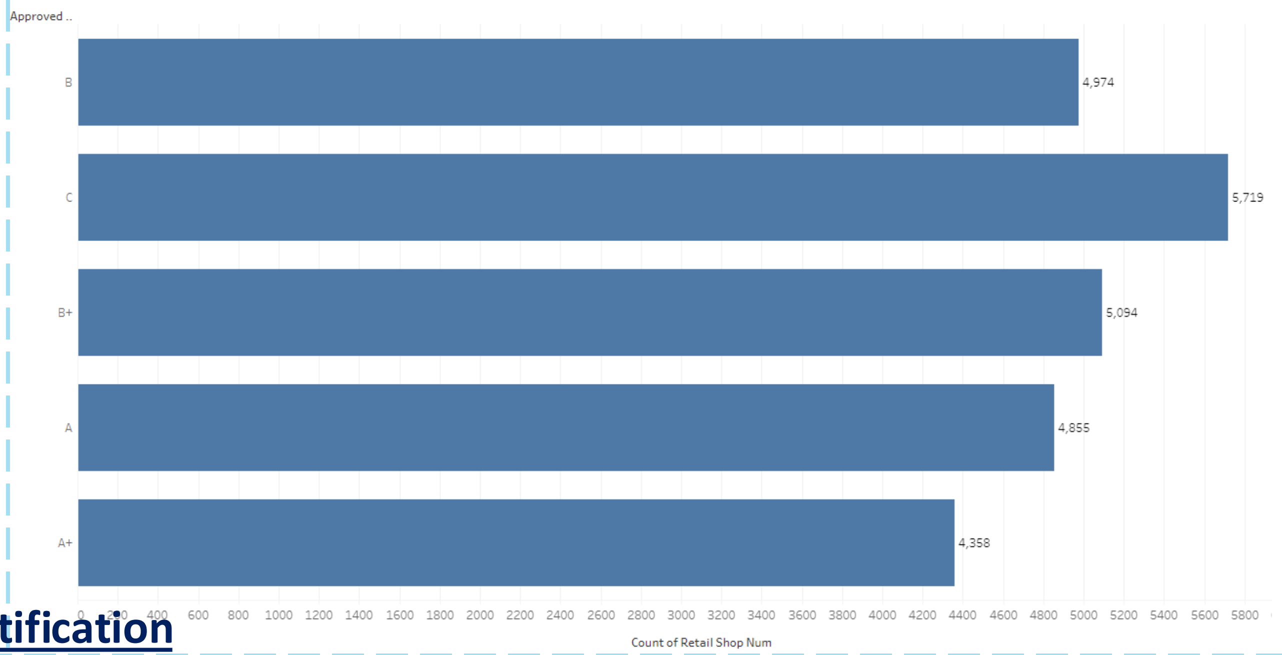

Storage issues
WH regiona.. Zone
Zone 1 East
North 13,371
South 10,451
West 7,407
Zone 2 North 13,305
South 23,207
West 10,543
Zone 3 East
North 14,829
South 13,267
West 14,652
Zone 4 East
North 18,050
South 13,556
West 3,504 30,450
Zone 5 East
North 3,221 29,988
South 17,028
West 22,538
Zone 6 East
North 7,926 69,688
South 21,508
West 3,346 37,589
0K 5K 10K 15K 20K 25K 30K 35K 40K 45K 50K 55K 60K 65K 70K 75K
Storage Issue Reported L3M


Breakdowns
WH regiona.. Zone
Zone 1 East
North 2,719
South 2,155
West 1,550
Zone 2 North 2,686
South 372 4,687
West 2,049
Zone 3 East 357
North 3,080
South 2,764
West 2,992
Zone 4 East 460
North 442 3,593
South 2,795
West 665 6,354
Zone 5 East
North 579 6,103
South 3,535
West 466 4,719
Zone 6 East
North 1,453 14,162
South 4,320
West 659 7,733
0K 1K 2K 3K 4K 5K 6K 7K 8K 9K 10K 11K 12K 13K 14K 15K 16K
Wh Breakdown L3M

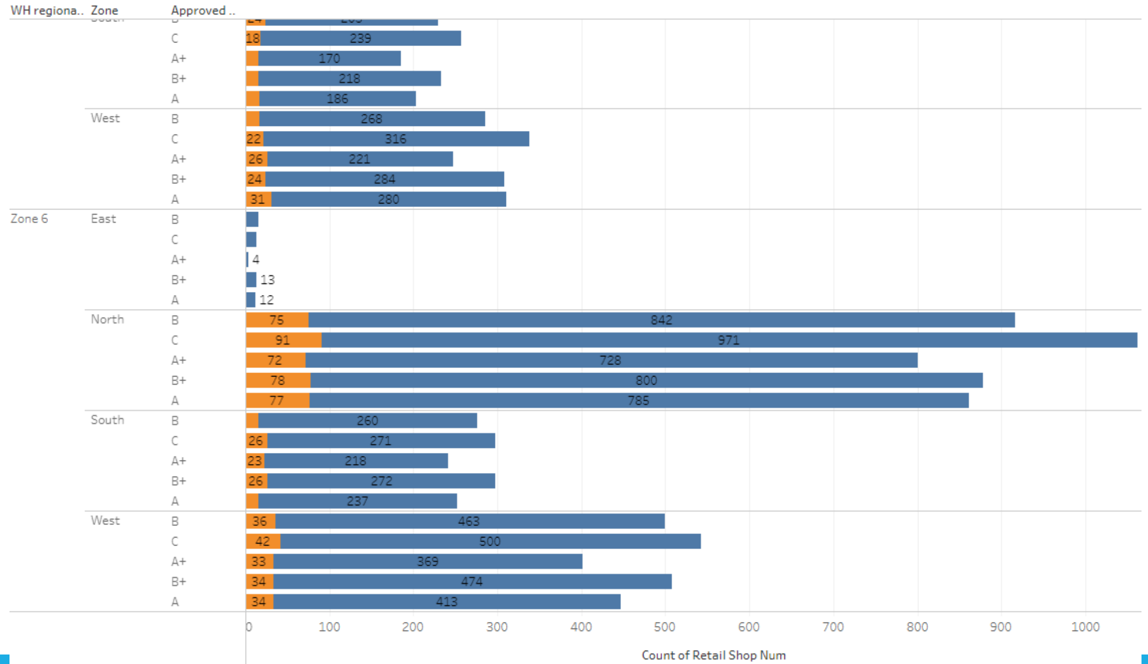

No. of Govt. checks
WH regiona.. Zone
Zone 1 East
North 15,147
South 13,342
West 9,080
Zone 2 North 9,114
South 33,506
West 6,081
Zone 3 East
North 22,024
South 17,522
West 17,050
Zone 4 East 3,432
North 27,872
South 18,720
West 3,114 32,522
Zone 5 East
North 32,196
South 17,562
West 19,060
Zone 6 East
North 7,659 80,775
South 18,591
West 32,565
0K 5K 10K 15K 20K 25K 30K 35K 40K 45K 50K 55K 60K 65K 70K 75K 80K 85K
Govt Check L3M

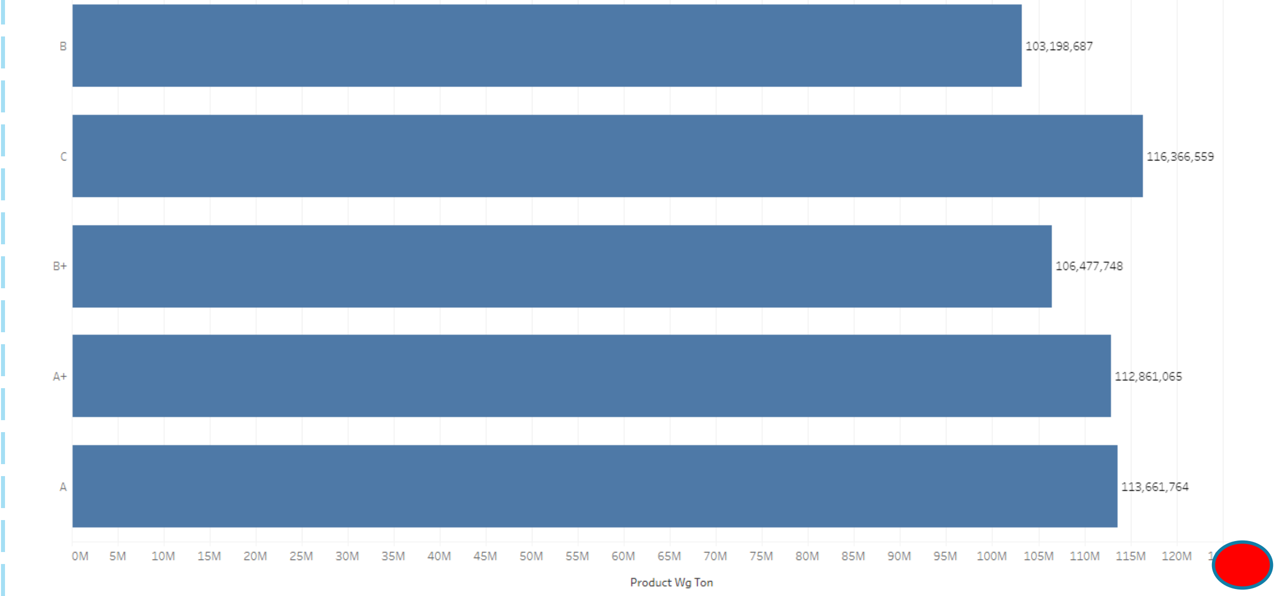
No. of WHs with Electric Backup

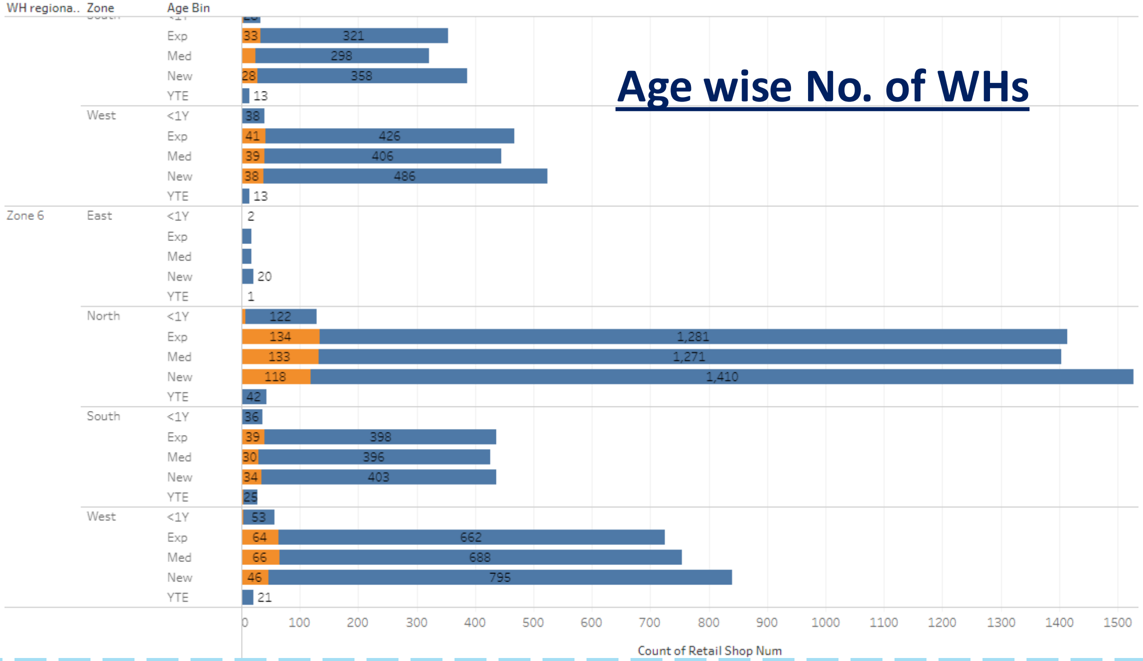




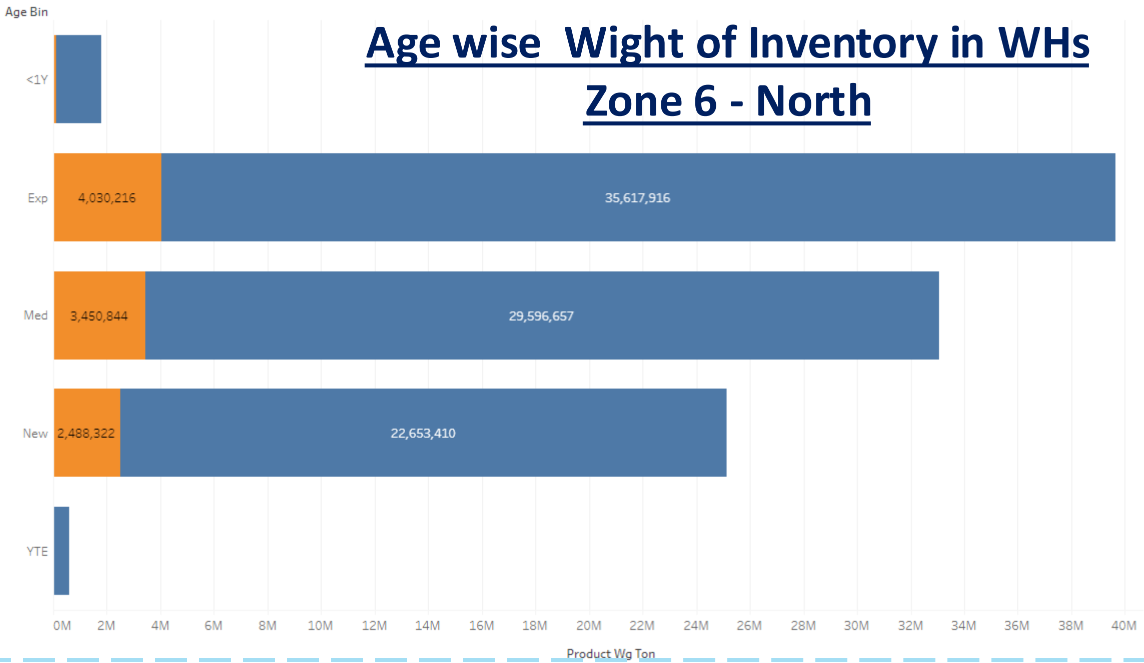
WH Certification Category

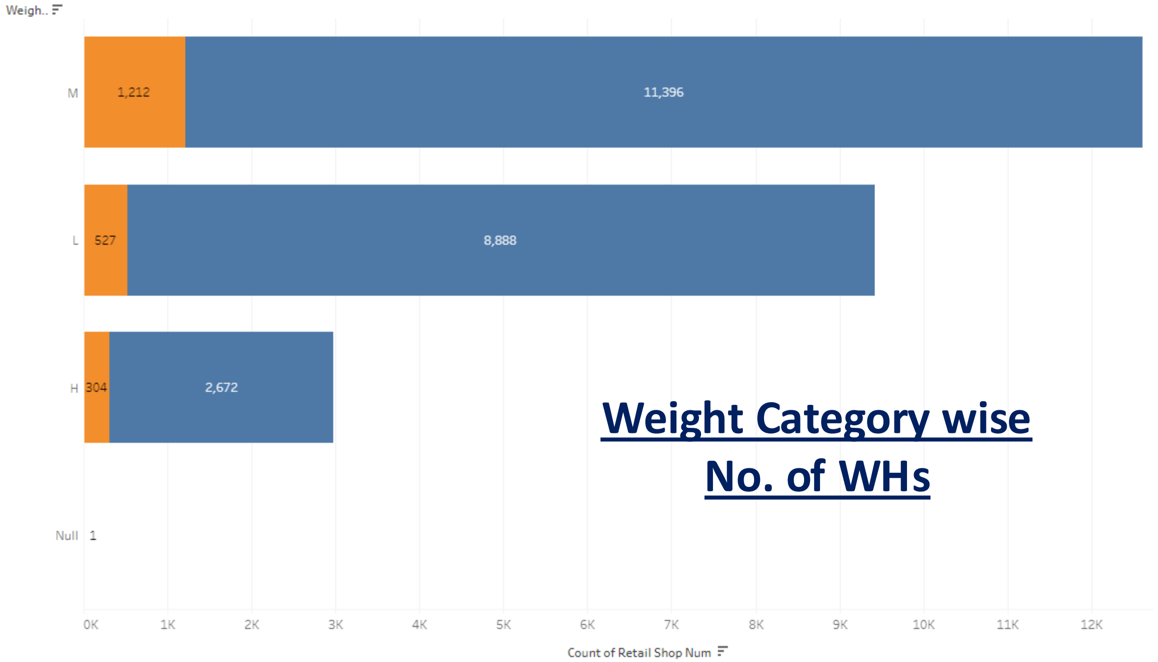


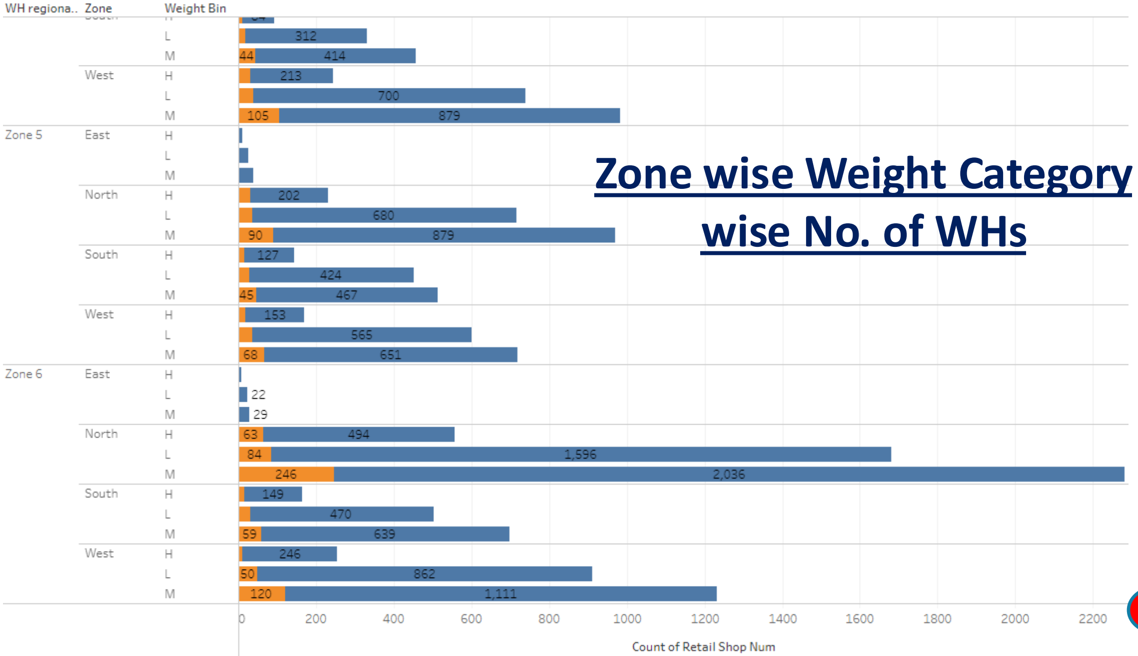




Location type
Rural
Urban







Insights from EDA

- ✓ Multi collinearity present in the data – Correlation Plot
- ✓ Warehouses in the Northern region & rural location type of Zone 6 are the most prominent
- ✓ Most of the Warehouses are C certified & carry medium category weight
- ✓ Most of the Ware houses are new but the customer base is strong for expert category Warehouses

Recommendations from EDA

- ✓ Analysis of market requirement
- ✓ Planning of warehouse inventory
- ✓ Timely Introduction of new products
- ✓ Targeted marketing & Strategic Pricing
- ✓ C to A+ certifications conversions

FEATURE TRANSFORMATION & SELECTION

- ✓ Encoding of Object type variables – One Hot Encoding & Dummy Encoding
- ✓ Multi collinearity treatment through Variance Inflation Factor method
- ✓ Measure of how much the variable is contributing to the standard error
- ✓ A value of up to 5 for VIF is allowed.
- ✓ The variables with a VIF above 5 are removed one by one and the VIF is recalculated after each removal.

	variables	VIF
8	govt_check_l3m	4.899737
7	approved_wh_govt_certificate	4.807978
5	storage_issue_reported_l3m	3.800284
0	num_refill_req_l3m	3.485383
4	electric_supply	2.877205
10	wh_owner_type_Rented	1.848336
12	WH_regional_zone_Zone 6	1.735769
6	temp_reg_mach	1.629826
14	WH_capacity_size_Small	1.417972
1	transport_issue_l1y	1.402753
13	zone_South	1.397069
11	WH_regional_zone_Zone 5	1.383800
2	flood_impacted	1.153500
9	Location_type_Urban	1.096589
3	flood_proof	1.080488

14 FINAL SIGNIFICANT VARIABLES

Model Building

Model Description

- ✓ Supervised Regression Problem
- ✓ Predict any value between 0 to + infinity for weight

Models for Regression Problems

- ✓ Linear Regression
- ✓ Decision Tree Regressor - CART
- ✓ Ensemble models
  - ❖ Random Forest Regressor
  - ❖ Adaboost
  - ❖ Bagging
  - ❖ Gradient boost Regressor

Train Test Split

- ✓ Train the model using train data and test using test data
- ✓ Separation of Dependent & Independent variables.
- ✓ Split data in to Train & Test data (70:30)

LINEAR REGRESSION

Model Definition & Building Approach

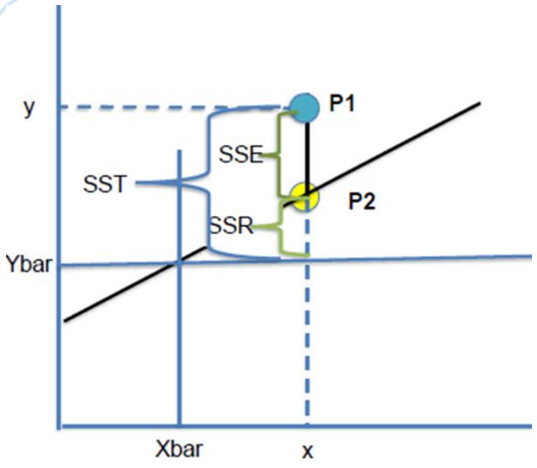
- ✓ Linear combination of the explanatory variables
- ✓ An expression of one or more variables scaled by a constant factor and added together
- ✓ Dependent variable value = (weight ∗independent variable) + constant
- ✓ It is the straight line in the scatter plot of the variables
- ✓ The best fit line can be found out using the Gradient Descent method
- ✓ Linear Regression models are built using SKLEARN & SCIPY STATS in Python

$$Y = m1X1 + m2X2 + \ldots\ldots + mnXn + C + e$$

Predictive power

Residual error

Errors & Metrics in linear regression model



P1	Original y data point for given x
P2	Estimated y value for given x
Ybar	Average of all Y values in data set
SST	Sum of Square error Total (SST), Variance of P1 from Ybar $(Y\ Ybar)^2$
SSR	Regression error (p2 ybar) 2 (portion SST captured by regression model)
SSE	Residual error $(p1\ p2)^2$

Best Fit line, SSE = 0 => SSR/ SST =1

R^2 = SSR/SST (Coefficient of Determination)

R^2 tends to 1

Lower the RMSE better the model

Train & Test model

- ✓ The model is initialized using the specific library
- ✓ Train data is fit in to the model & model gets trained for predictions
- ✓ The model is tested using test data and the metrics R^2 & RMSE are evaluated

DECISION TREE REGRESSOR

Model Definition & Building Approach

- ✓ CART – Classification And Regression Tree is a binary decision tree
- ✓ Gives both categorical & continuous output
- ✓ Nodes are split based on the least mean squared error for regression
- ✓ Model is built using the SKLEARN library in Python
- ✓ Errors, Metrics, training & testing models remain the same as in linear regression model.

Hyper Parameters

- ✓ Decision Tree Regressor function is called and initialized with certain hyper parameters
- ✓ Hyper parameters determines the way the nodes are split and their criteria
- ✓ Criterion – Squared error is the split criterion (for regression models)
- ✓Maximum depth - number of branches that the tree can be split along vertically
- ✓ Min sample leaf - minimum number of records that a node must contain after splitting
- ✓ Min sample split - minimum number of records that a node must have so that it can be split
- ✓ The model is built for default parameters & further tuned to avoid over fitting (Exceptional in train data, while fails in test data)

Model Tuning

- ✓ Grid search cross validation-
- ✓ Grid search - technique to select the best ML model, parameterized by a grid of hyper parameters
- ✓Cross validation – splits dataset into random groups, holding one group as test, and training the model on the remaining groups
- ✓Process is repeated for each group being held as test, then the average of models is used for the resulting model.

RANDOM FOREST REGRESSOR

Model Definition, Building Approach & Tuning

- ✓ Ensemble technique - multiple CART models are built to get the accuracy in the predicted value
- ✓ Model building is same except that we need to build a Random Forest Regressor model instead of Decision Tree Regressor
- ✓ Errors, Metrics, training & testing models remain the same as in the other two models
- ✓ Model tuning is same as in that of the Decision tree regressor

Hyper Parameters

- ✓ Same as that of DT Regressor
- ✓ Some additional hyper parameter include Max_features (no. of independent variables to be considered) and n_estimators (no. of DTs to be built)
- ✓ The model is built for default parameters & further tuned to avoid over fitting

BOOTSTRAP AGGREGATING

Model Definition & Building Approach

- ✓ A technique of sampling - creates sub sample of observation with the actual dataset, with replacementme
- ✓ Reduced chances of over fitting
- ✓ Self imposed Model tuning - trains a large number of “strong” learners in parallel & combines to "smooth out" predictions
- ✓ Model is built using the Bagging Regressor of SKLEARN in python
- ✓ Model initialization includes defining the model over which the Bagging has to happen. In this case it is the RF Regressor

BOOSTING

- ✓ Linear sequential process - model minimizes errors by learning from previous model predictions
- ✓ Large number of weak learners trained sequentially – Combines to forma a strong learner
- ✓ Two types of boosting – Adaptive/ Ada boost & Gradient boosting

Ada Boost

- ✓ Successive learners are created with a focus on the ill fitted data of the previous learner
- ✓ Adaboost Regressor library of Sklearn is used to build the model
- ✓ Initialized with number of Decision trees

Gradient Boost

- ✓ Each learner is fit on a modified version of original data
- ✓ New models are fit to the residuals
- ✓ The overall learner improved where the residuals are high
- ✓ Gradient boosting Regressor library of Sklearn is used to build the model

Model Results & Insights

Model	Metrics				Model Fitness	Business Implication
	Train Data		Test Data			
	R2	RMSE	R2	RMSE		
Linear Regression using SKLEARN	0.9766	1780.4493	0.9777	1716.9875	Good	Larger value of RMSE shows that variation in prediction with rest to actual weight is high and predictions might sometime lead to Inventory cost loss in the ware house.
Linear Regression using SKLEARN - Scaled	0.9766	0.1528	0.9777	0.149	Good	
Linear Regression using Statsmodel	0.977	1780.4493	0.977	1716.9875	Good	
Linear Regression using Statsmodel after removing non significant variables	0.977	1780.72	0.977	1716.7562	Good	
CART Model (DT)	0.9998	154.7	0.9878	1270.2599	Poor	Model performs exceptionally well in the train data with a low RMSE amongst the models built. However, the model is overfit that is evident from the wide gap between the RMSEs of train and test data. Deploying this model for production would incur huge losses to the business
Reg. Cart DT (max_depth =10, min_samples_leaf=10, min_samples_split=30)	0.9941	891.524	0.9935	926.8318	Good	Larger value of RMSE shows that variation in prediction with rest to actual weight is high and predictions might sometime lead to Inventory cost loss in the ware house.
Reg. Cart DT Grid Search CV (max_depth =10, min_samples_leaf=10, min_samples_split=45)	0.9939	904.7152	0.9935	925.018	Good	
Reg. Cart DT Grid Search CV (max_depth =9, min_samples_leaf=5, min_samples_split=15)	0.9939	908.675	0.9935	922.737	Good	
Reg. Cart DT Grid Search CV (max_depth =9, min_samples_leaf=5, min_samples_split=12)	0.9939	908.1311	0.9935	923.33	Good	
Random Forest Model	0.9988	390.449	0.993	963.7913	Poor	Model performs exceptionally well in the train data with a low RMSE amongst the models built. However, the model is overfit that is evident from the wide gap between the RMSEs of train and test data. Deploying this model for production would incur huge losses to the business
Random Forest Model with Grid Search CV (max_depth = 10, max_features=7, min_samples_leaf=5, min_samples_split=15, n_estimators=301)	0.9927	991.32	0.9919	1031.8729	Good	Larger value of RMSE shows that variation in prediction with rest to actual weight is high and predictions might sometime lead to Inventory cost loss in the ware house.
Random Forest Model with Grid Search CV (max_depth = 10, max_features=9, min_samples_leaf=10, min_samples_split=30, n_estimators=150)	0.9938	913.4476	0.9935	924.6953	Good	
Bagging	0.9974	587.0825	0.9934	929.7687	Poor	Model performs exceptionally well in the train data with a low RMSE amongst the models built. However, the model is overfit that is evident from the wide gap between the RMSEs of train and test data. Deploying this model for production would incur huge losses to the business
AdaBoost	0.9986	427.6	0.993	958.0611	Poor	
Gradient Boosting	0.9935	936.6518	0.9938	902.504	Good	Larger value of RMSE shows that variation in prediction with rest to actual weight is high and predictions might sometime lead to Inventory cost loss in the ware house.

- ✓ Gradient boosting has the least RMSE in test data & RMSE has decreased for testing data
- ✓ Implies the least possible inventory cost loss
- ✓ Aids in pre-planning and specific campaigns
- ✓ Boost sales which aids to further boosts the bottom line of company.

BUSINESS RECOMMENDATIONS

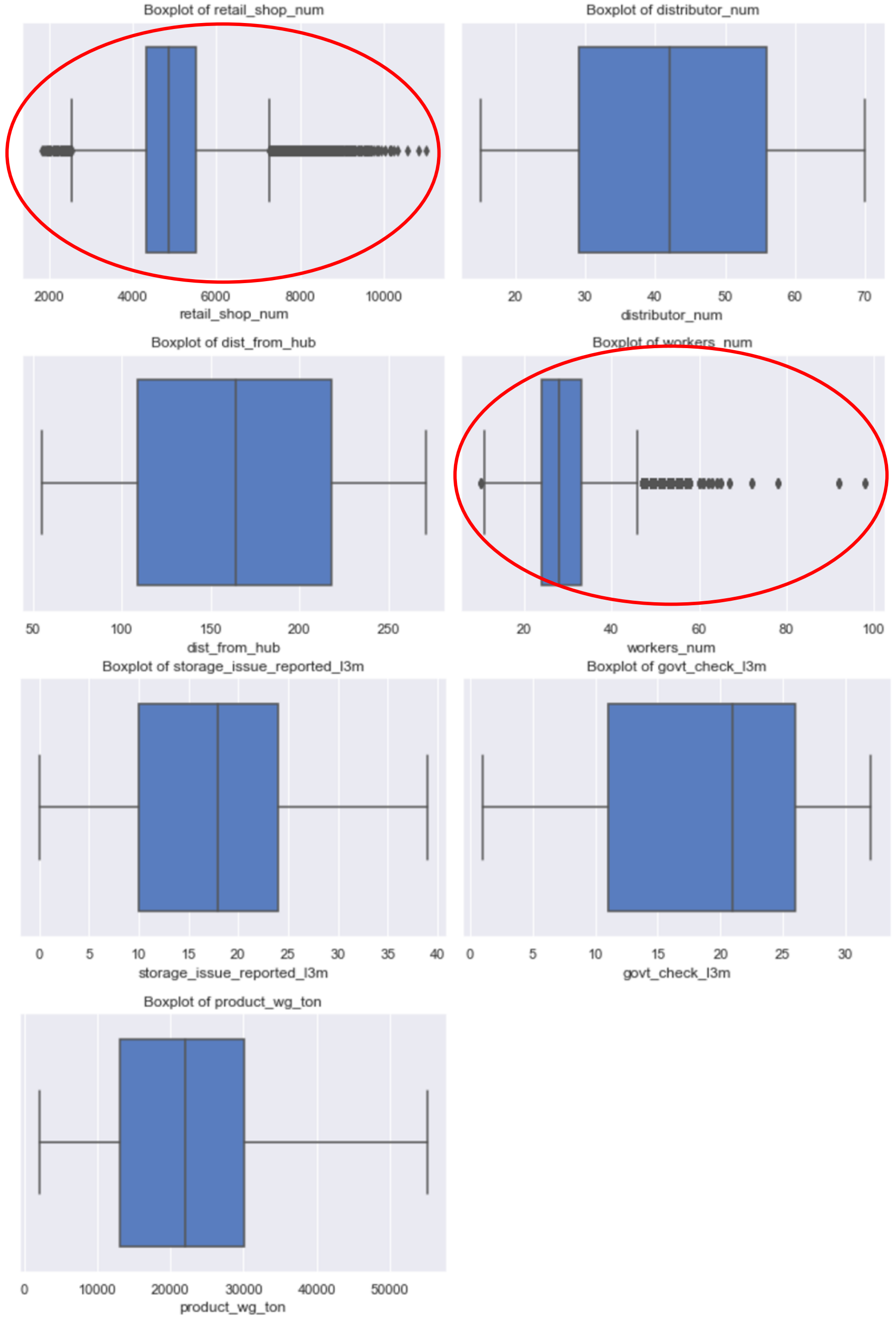
- ✓ Gradient boosting can be deployed for predictions as it has the least RMSE in test data & RMSE has decreased for testing data
- ✓ Analysis of market requirement in Northern region, rural location of Zone 6
- ✓ Proper planning of warehouse inventory
- ✓ Timely introduction of new products
- ✓ Targeted marketing based on demography & Strategic Pricing
- ✓ C to A+ certifications conversions of all the warehouses in turn developing the customer base

Thank You

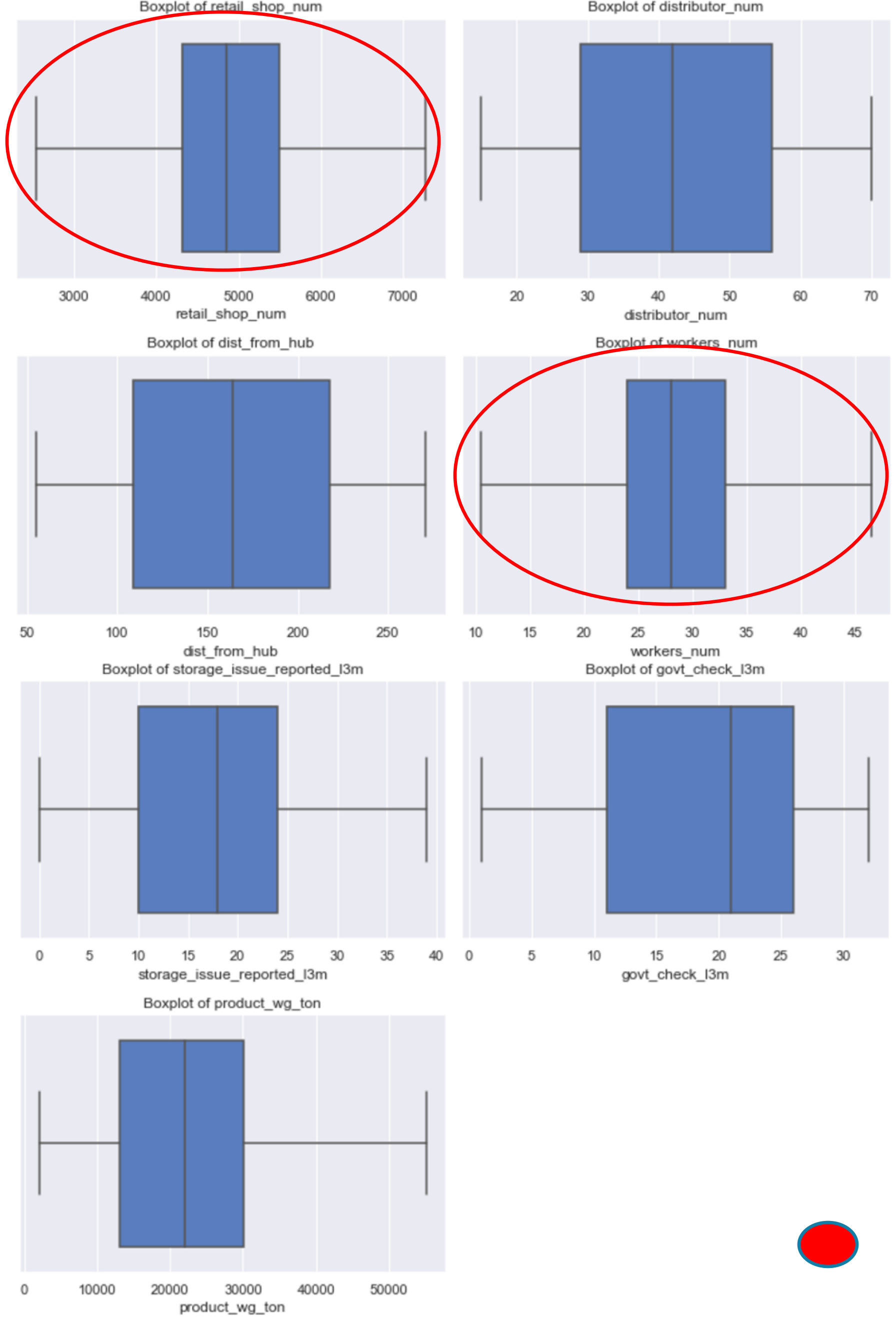
APPENDIX

Appendix 1

BEFORE



AFTER



Appendix 2

Age Bins	Description
-1	Yet to Establish
0	Less than 1 year
1 to 9	New
10 to 17	Mediocre
18 to 26	Expert

Weight Bins	Description
2065 to 17695.33	Low
17695.34 to 35390.67	Medium
35390.68 to 55151	High