Scope

Introduction

The paradigm of favoring data driven outcomes over conventional methods in making critical business decisions is disruptive to the traditional business practice. Although unconventional, the acute nature of competition and critical resource constraints continues to pave the way for such considerations.

A large logistic company faces a complex challenge in effectively managing the billing process whenever a customer misuses their shipping labels. XYZ LOGISTICS billing directly depends on the shipping information generated by the customer. The PLD creation process also generates the shipping label. If a customer makes multiple photo copy of the shipping label (that was generated during the single shipment creation) and uses them to ship multiple packages, XYZ LOGISTICS would not be able to bill all the subsequent packages (without manual intervention). However, XYZ LOGISTICS will provide service on all the subsequent packages.

It is a manually intensive process to analyze the scans on each package to identify the possible duplicate movement and bill them.

This project focuses on developing a functionally adequate logistic model with the core purpose of predicting how likely (or the probability of) a package movement being duplicated once or multiple times. Additionally, a second linear model has to be developed to determine the number of times the label was duplicated. In order to develop statistically sound and functional predictive models, it is important that the following items are taken into consideration.

- Identifying an optimal list of predictors so that the balance between the bias and variance of the model is optimized.
- Cross validating the model with out of sample data to ensure the quality of predictability.
- Assessing the goodness of fit for the model. The model should not be grossly violating the assumptions associated with the development of multiple linear regression models.

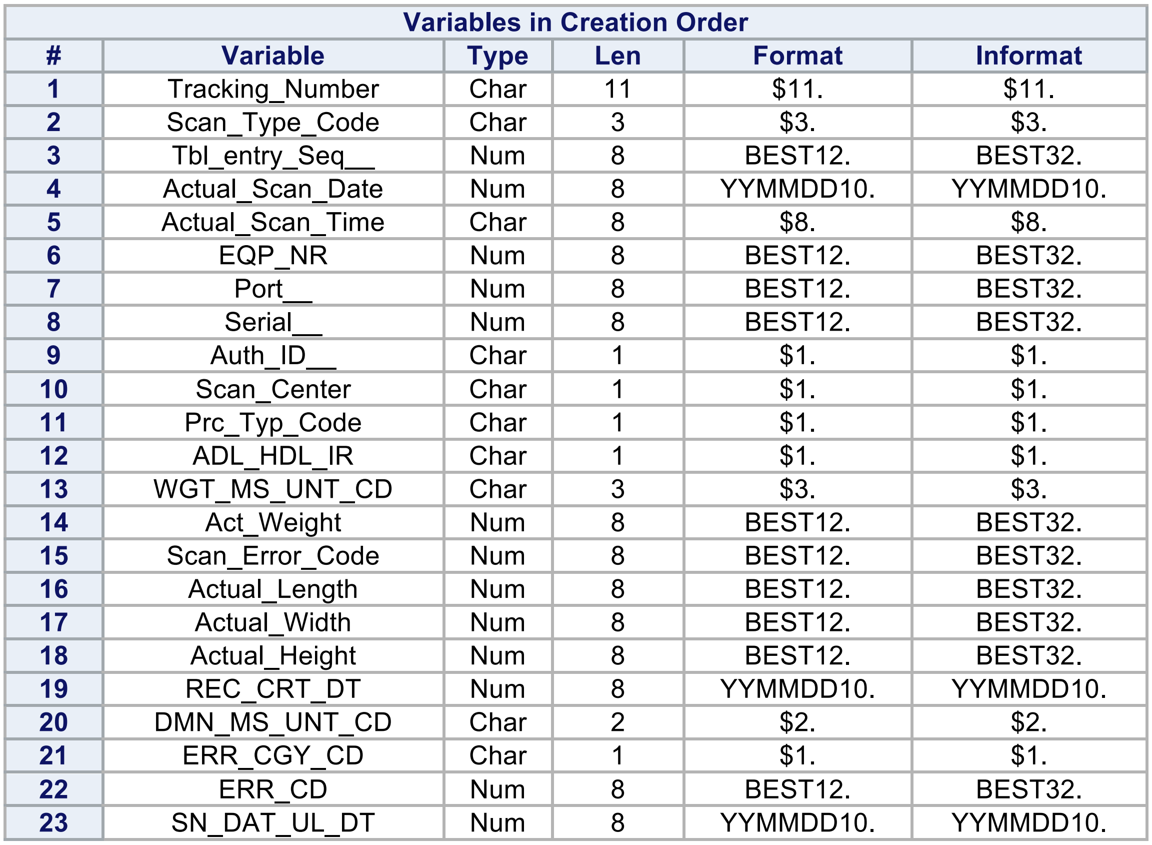
The core focus for the data analysis steps was concentrated on the following.

- Developing appropriate features based on aggregation of data
- Developing the logistic model and predicting the probability of a label getting duplicated.
- Developing a linear model to predict the number of times the label was duplicated.

Data

The following two sets of data have been used in developing the model.

- **DWS Scan Data** Dimension Scan Data.



Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	Tracking_Number	Char	11	$11.	$11.
2	Scan_Type_Code	Char	3	$3.	$3.
3	Tbl_entry_Seq__	Num	8	BEST12.	BEST32.
4	Actual_Scan_Date	Num	8	YYMMDD10.	YYMMDD10.
5	Actual_Scan_Time	Char	8	$8.	$8.
6	EQP_NR	Num	8	BEST12.	BEST32.
7	Port__	Num	8	BEST12.	BEST32.
8	Serial__	Num	8	BEST12.	BEST32.
9	Auth_ID__	Char	1	$1.	$1.
10	Scan_Center	Char	1	$1.	$1.
11	Prc_Typ_Code	Char	1	$1.	$1.
12	ADL_HDL_IR	Char	1	$1.	$1.
13	WGT_MS_UNT_CD	Char	3	$3.	$3.
14	Act_Weight	Num	8	BEST12.	BEST32.
15	Scan_Error_Code	Num	8	BEST12.	BEST32.
16	Actual_Length	Num	8	BEST12.	BEST32.
17	Actual_Width	Num	8	BEST12.	BEST32.
18	Actual_Height	Num	8	BEST12.	BEST32.
19	REC_CRT_DT	Num	8	YYMMDD10.	YYMMDD10.
20	DMN_MS_UNT_CD	Char	2	$2.	$2.
21	ERR_CGY_CD	Char	1	$1.	$1.
22	ERR_CD	Num	8	BEST12.	BEST32.
23	SN_DAT_UL_DT	Num	8	YYMMDD10.	YYMMDD10.

Fig. 1: SAS PROC CONTENTS: DWS SCAN Data

- **SPA Data.** To be analyzed

The DWS Scan data set (for a particular week) contained 891755 observations. Each observation corresponds to a scan event for a particular shipping label (identified by tracking number). A total of 315885 unique tracking numbers were present in this dataset

Data Survey and Data Quality Check.

Although, the dataset contained multiple fields with missing values, the items of interest were comparatively clean. The following variables (in the dataset) are specific items of interest.

- Tracking_Number
- Actual_Scan_Date
- Actual_Scan_Time
- Serial__
- Act_Weight

- Actual_Length
- Actual_Width
- Actual_Height

Since the Actual Weight was missing during multiple scans events, the weight was computed from the length, breadth and width field using the standard XYZ LOGISTICS weight calculation formula (l*b*h/166). If the size dimensions were missing, the actual weight was used to impute the weight field. If both the dimensions and the actual weight were missing, the records were ignored.

No other major issues were found with this dataset in terms of missing values of irregularities with the data collection process. Hence, this dataset has been qualified for the purpose of EDA and development of critical features.

Outlier Removal.

Based on EDA of the raw training data (prior to any transformations), the following actions were taken to deal with possible outliers that may be present in the data.

- TBD.

Analysis

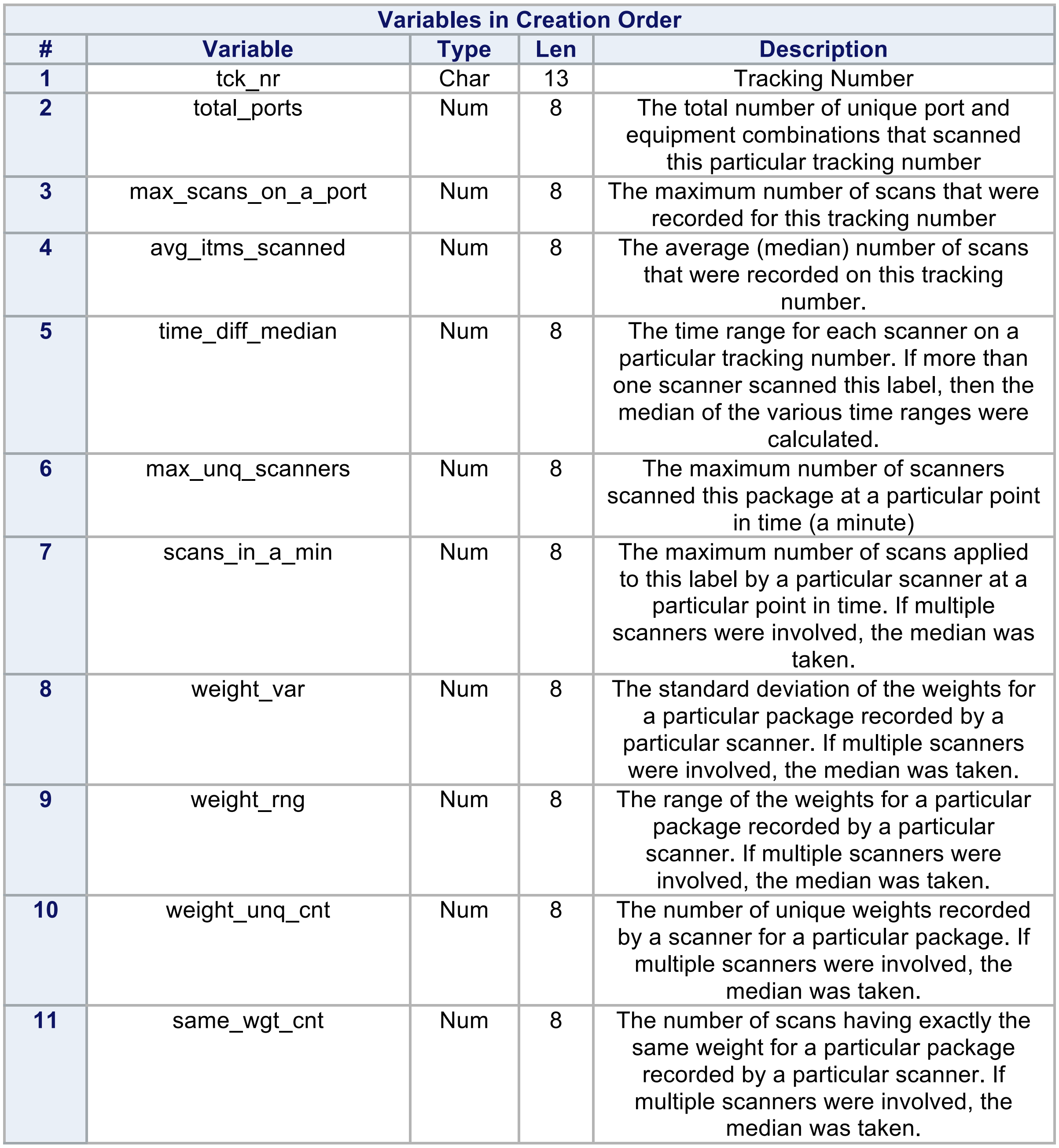
The next stage of this report will focus on the following deliverables.

- Generation of high value features and exploration of the data.
- Determining a strategy to train a supervised model
- Identifying a list of highly relevant predictors and developing the logistics models.
- Assessing the transformation requirements for each of the variables (if applicable).
- Apply various methods of automatic variable selection methods in defining the optimal model.
- Cross-validate the model against the remaining 30% of the data and analyze the quality of prediction.
- Selecting the best model based on ROC AUC, AIC, Out of Sample Confusion Matrix

A final conclusion will be provided to decide the most appropriate model and the next steps.

Developing Critical Features.

The following features were developed from the DWS Scan data



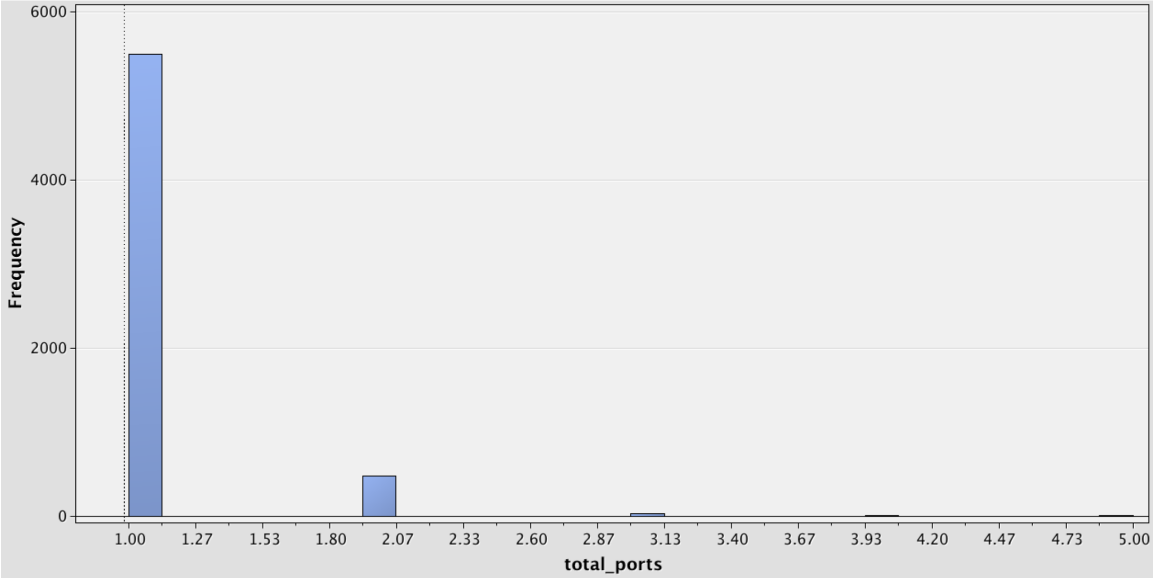
Variables in Creation Order

#	Variable	Type	Len	Description
1	tck_nr	Char	13	Tracking Number
2	total_ports	Num	8	The total number of unique port and equipment combinations that scanned this particular tracking number
3	max_scans_on_a_port	Num	8	The maximum number of scans that were recorded for this tracking number
4	avg_itms_scanned	Num	8	The average (median) number of scans that were recorded on this tracking number.
5	time_diff_median	Num	8	The time range for each scanner on a particular tracking number. If more than one scanner scanned this label, then the median of the various time ranges were calculated.
6	max_unq_scanners	Num	8	The maximum number of scanners scanned this package at a particular point in time (a minute)
7	scans_in_a_min	Num	8	The maximum number of scans applied to this label by a particular scanner at a particular point in time. If multiple scanners were involved, the median was taken.
8	weight_var	Num	8	The standard deviation of the weights for a particular package recorded by a particular scanner. If multiple scanners were involved, the median was taken.
9	weight_rng	Num	8	The range of the weights for a particular package recorded by a particular scanner. If multiple scanners were involved, the median was taken.
10	weight_unq_cnt	Num	8	The number of unique weights recorded by a scanner for a particular package. If multiple scanners were involved, the median was taken.
11	same_wgt_cnt	Num	8	The number of scans having exactly the same weight for a particular package recorded by a particular scanner. If multiple scanners were involved, the median was taken.

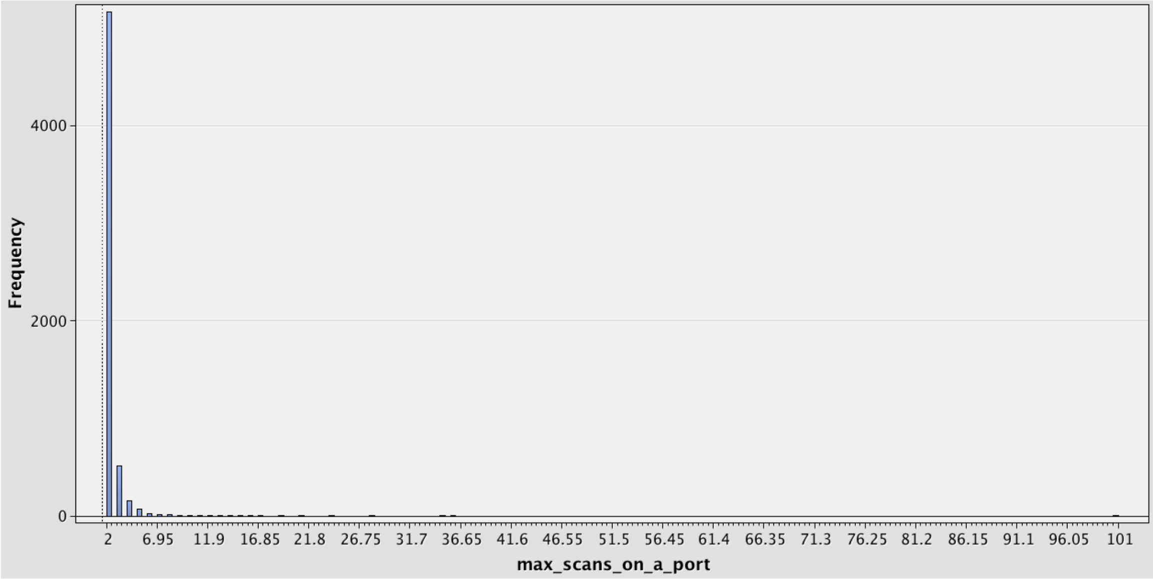
An initial EDA on the above features shows the following.

Variable	Minimum	5th Pctl	Mean	Median	75th Pctl	95th Pctl	Maximum	Std Dev	N Miss
total_ports	1.00	1.00	1.13	1.00	1.00	2.00	64.00	0.70	0
max_scans_on_a_port	1.00	2.00	2.44	2.00	2.00	4.00	226.00	1.77	0
avg_itms_scanned	1.00	2.00	2.39	2.00	2.00	4.00	163.50	1.56	0
time_diff_median	0.00	120.00	8298.58	1260.00	5100.00	41460.00	602220.00	22543.19	0
max_unq_scanners	1.00	1.00	1.07	1.00	1.00	2.00	44.00	0.33	0
scans_in_a_min	1.00	1.00	1.02	1.00	1.00	1.00	22.00	0.16	0
weight_var	0.00	0.00	0.43	0.07	0.15	0.85	433.67	3.34	1
weight_rng	0.00	0.00	0.66	0.10	0.30	1.30	613.30	5.27	0
weight_unq_cnt	1.00	1.00	1.70	2.00	2.00	3.00	48.00	0.68	0
same_wgt_cnt	1.00	1.00	1.65	1.50	2.00	3.00	133.00	1.19	0

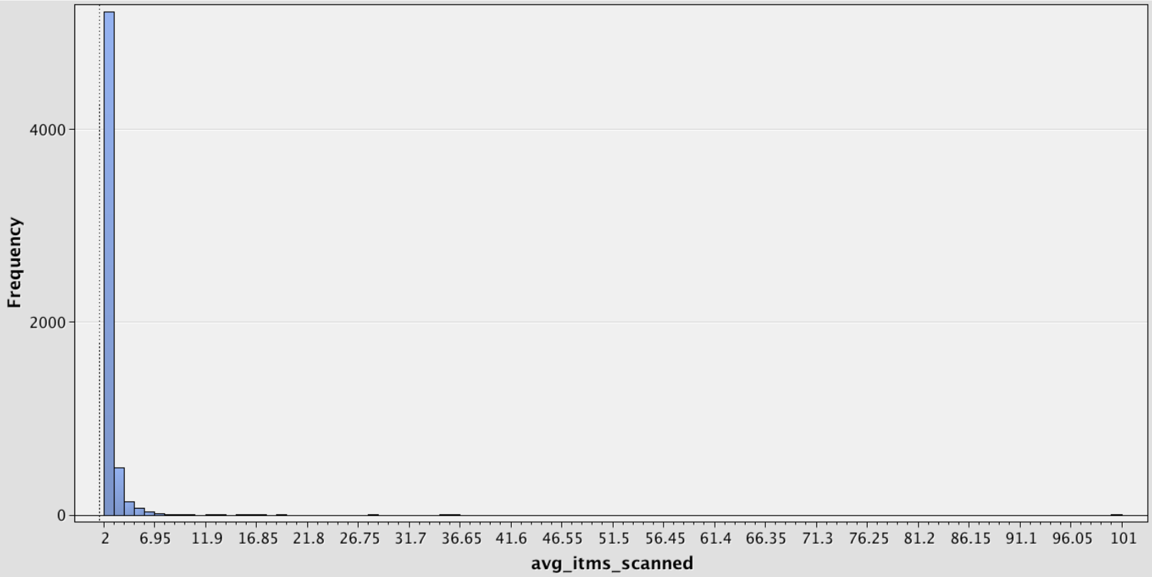
- total_ports – 1012 out of 315885 had been scanned by more than 3 ports



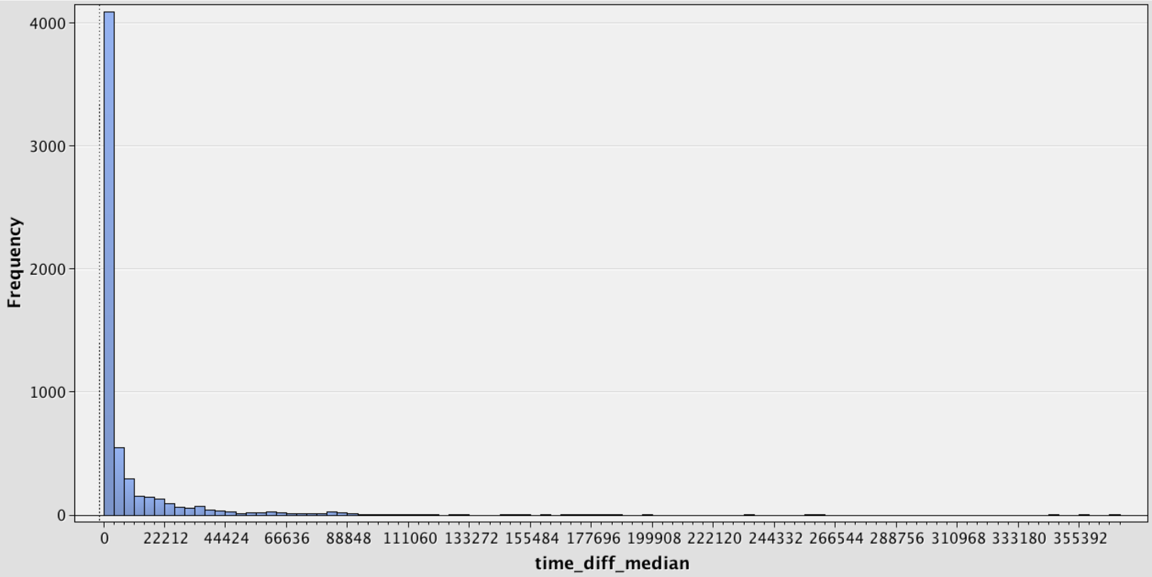
- Max_scans_on_a_port - 14039 out of 315885 had minimum of 5 or more scans by a particular scanner.



- avg_itms_scanned - 12591 out of 315885 had an average of 5 or more scans by a particular scanner.



- time_diff_median – 92871 out of 315885 experienced a scan period of more than an hour by a particular scanner



- max_unq_scanners – 581 out of 315885 experienced a scan by more than 3 scanners at a particular point in time (within a minute).

- scans_in_a_min – 4901 out of 315885 experienced 3 or more scans by any particular scanner within a minute.

- weight_var – 13983 out of 315885 had a weight standard deviation of more than 1 lb based on weights measured by any particular scanner.

- weight_rng - 8733 out of 315885 had a weight difference of more than 3 lb based on minimum and maximum weights measured by any particular scanner.

- weight_unq_cnt – 4541 out of 315885 had a measure of 3 or more unique weights as measured by any particular scanner.

- same_wgt_cnt – 9947 out of 315885 had had exactly the same weight recorded on 3 or more unique scans as measured by any particular scanner.