**Initial Analysis of the data**

I) **Changes made to the excel is attached.**

****

II) **Mean Square Error for each Distributor - :**

**The sheet shows the difference of the testing data with the prediction data for the complete data set.**

****

*We can draw a table for the monthly data putting up a filter for the month as well as of the distributor’s code.*

Total Mean Square error (no filter) -: 91385.35

The mean square error is the sum of squared difference on each prediction to the actual value.

For Example the Complete distribution has the MSE as 91385.35 and total number of rows as 63719.

So each row has squared error as 91385.35 / 63719, which are approx. 1.43 Hence on the average the prediction of the actual cost is +/- Square root (1.43)

For a distributor, let’s say SS Drinks here is the prediction data, with total MSE = 16048.5



III)Model Summary:-

The accuracy of the model is currently 79% as we have taken Linear and Bayesian model for the regression purpose. The Features are not sufficient enough for the more accurate prediction.

As Distribution Code TDM and Distributor are all three same value Columns so it seems justified taking any one (we took the code as it’s an integer).

We removed the date for each month to train the model as per the monthly data and not on the data purpose as the number of features are low for that.

IV) Future Work-:

Also another perspective to work on the analysis of this dataset is using Time Series Model. It sounds feasible and worth a good try to have around 3-4 years of monthly data and fit it with a time series model likes ARMA. It could be the case that the monthly data is dependent on the previous month or previous couple of month (might have a correlation).

Statistical Analysis is still pending of the given model due to time constraint like, which feature is more important that other and correlation among the features. Also see if there is any distribution the data is following.