

## Group 9

Akshat Chowdhary - 61310155

Anushree Gandhi - 61310245

Kashish Goyal -61310267

Ravdeep Chawla -61310227

Vaibhav Tripathi -61310069

# Forecasting Daily Sales of Perishable Foods to reduce spoilage

---

Leveraged data for the past year to generate a daily forecast of highly perishable vegetables for the next 30 days to reduce spoilage of exotic fruits and beans in the inventory. The biggest factors affecting sale are day of week and holidays. Using these two parameters, we can reasonably estimate sales in the future.

# Table of Contents

---

|  |    |
|--|----|
| Executive Summary .....                                      | 2  |
| Business Goal .....  | 3  |
| Data Mining Goal .....                                       | 4  |
| Naïve Forecasts .....  | 5  |
| Multi-Step Forecast (Holt-Winters + Linear Regression) ..... | 6  |
| Key Learning and Recommendations .....                       | 11 |
| Exhibits .....   | 12 |

# Executive Summary

---

**Problem Description** - Representing the hypermarket, the objective of our forecasting is to reduce the spoilage of vegetables in the hypermarket by accurately forecasting sales on a daily basis. By using historical data of the last year we plan to forecast daily demand for two SKU's 'Exotic Vegetables' and 'Beans' under the vegetable sub-class.

Since the two selected vegetable SKU's had a profit margin of close to 25% we tried to model the costs of under-prediction or over-prediction as an important metric to assess the performance of our model

**Model Description** - Final model used for our analysis is comprised of two steps and is a combination of two forms of forecasting method. We ran the Holt-Winter's Model on the data following it by a Multiple Linear Regression using the forecasts and the relevant holidays which do not fall on a weekend. We have used a Holt Winter's with no trend in the initial step as the visualizations of the raw data do not highlight any particular trend. However we have captured weekly seasonality where we observed that sales were highest on a Sunday every week followed by Saturday and Wednesday.

**Model Performance** - Since we had considered the Naive forecasts as a benchmark for our prediction we have taken and compared the RMSE and MAE parameters for our model obtained by conditioning the cost factor into our model with the Naive forecast models.

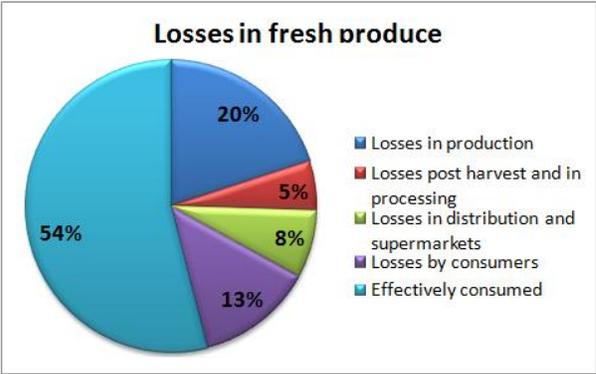
**Forecasts and their assumptions** - We generated one month forecast in the future to generate a daily forecast of vegetable supplies for the next 30 days. We planned to ensure that this model was updated every month by factoring in the errors that affected the model last month

**Conclusions and Recommendations** - We can fairly conclude that the model developed by us is effective to forecast daily vegetable sales for the next 30 days for the two vegetable SKU's. We observe that certain forecasts in July and Aug are not captured well. We believe this is the effect of a special event such as 'Rains' in the city disrupting normal events. We recommend that the predictive ability of the model can be improved by modeling in daily rainfall data.

# Business Goal

---

The business objective of the forecasting is to enable the Hypermarket to do weekly forecasts for next 30 days, so that they can avoid spillage. According to a report by the Food and Agriculture Organization of the UN, the lack of tuning between supply and demand is a big reason for spillage. In the distribution chain and supermarkets 7.5% fresh products are being lost due to degradation and expiring 'best-before-dates'. The chart breaks down the spillage by the stage at which they occur.



Not all losses are irreversible because a lot of effort has increasingly been put in the valorization of these products. Therefore, accurate prediction of demand is important in order to prevent such losses. Following stake holders are expected to benefit from this forecast.



# Data Mining Goal

---

The data mining goal is to generate a model that will predict monthly demand in advance, while accounting for following components:

- Level
- Trend
- Seasonality

Our model will use data for the past year to generate a daily forecast of vegetable supplies for the next 30 days, while updating this model every month to factor in the errors that affected the model last month.

The intended use of these forecasts would be where the hypermarket can send the daily level forecast to the suppliers on the first of every month. This would enable the supplier to truck the data efficiently on a daily basis. With this we expect to not only control the inventory costs for the hypermarket but also try and reduce the loss of revenue because of under prediction or over prediction.

This plays an important role specifically for the two chosen vegetables which are high margin products. According to our assumptions the two costs which would need to be factored into our data mining model apart from the inventory holding costs would be as below:

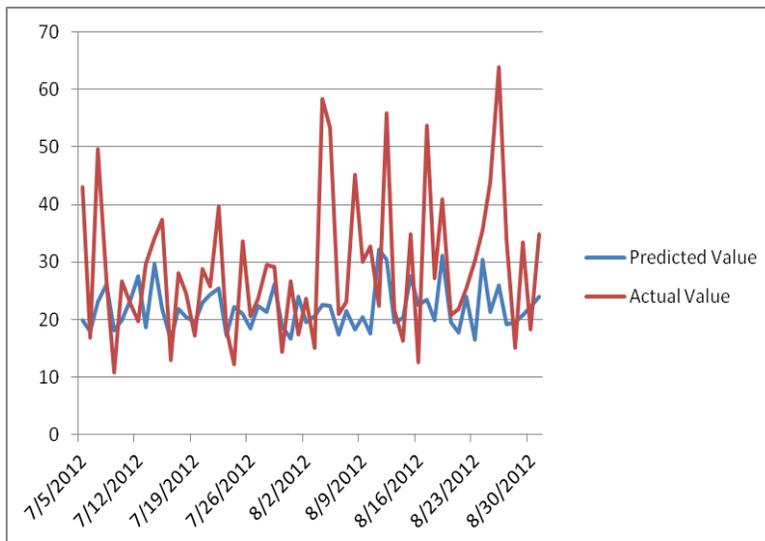
Cost of Over-stocking: 100% of Revenues generated from the absolute difference in actual and over stocked quantity

Cost of Under-stocking: 10% of Revenues generated from the difference in actual and under stocked quantity

# Naïve Forecast

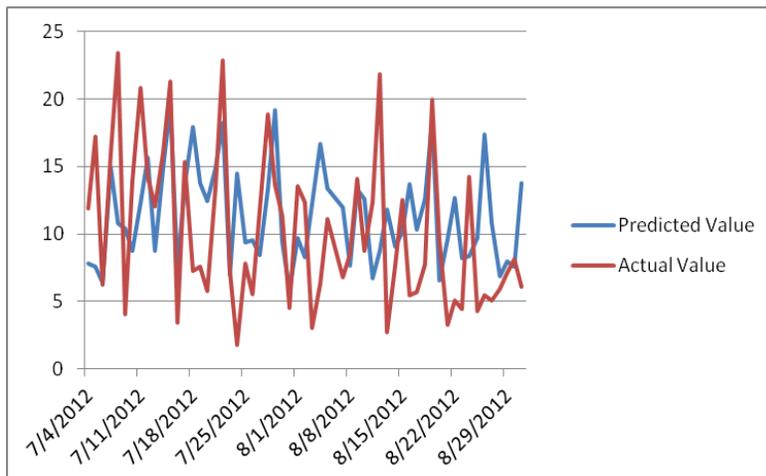
The Naive forecast for both 'Exotic Vegetables' and 'Beans' were modeled and we considered this forecast as a benchmark for evaluating the performance of our created model. The forecasts and the performance measures are as given below:

## Exotic Vegetables



|      |             |
|------|-------------|
| MAE  | 9.884988    |
| RMSE | 13.54212599 |

## Beans

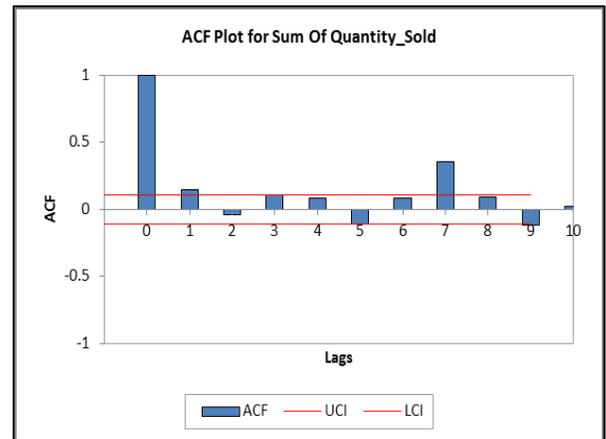
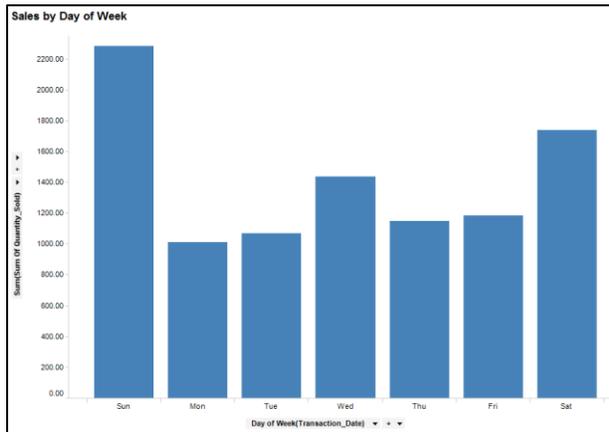


|      |             |
|------|-------------|
| MAE  | 4.789154    |
| RMSE | 5.912079212 |

# Multi-Step Forecast

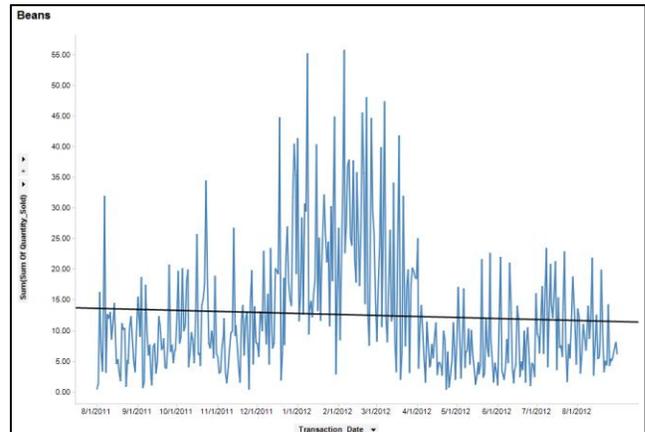
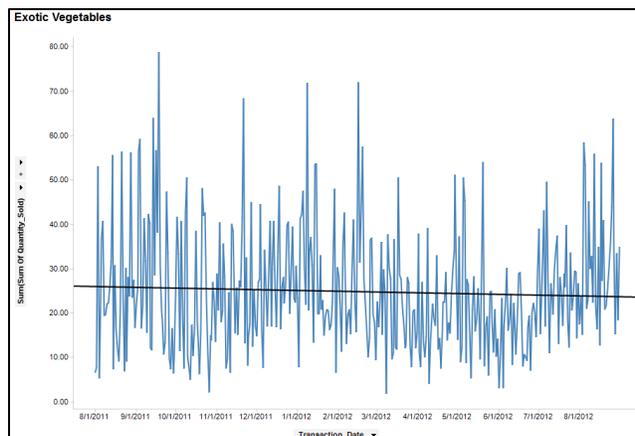
## SEASONALITY

The first observation we made on the data was that there was that sales were heavily dependent on the day of the week and followed a weekly seasonality. As shown in the two graphs below, there is a clear tendency of consumers to buy maximum on a Sunday followed by Saturday and then, in the middle of the week on Wednesday. Secondly, sales on time  $t$  are heavily correlated to time  $t-7$  days. The graphs are for exotic vegetables below but the trend is exactly similar for beans.



## TREND

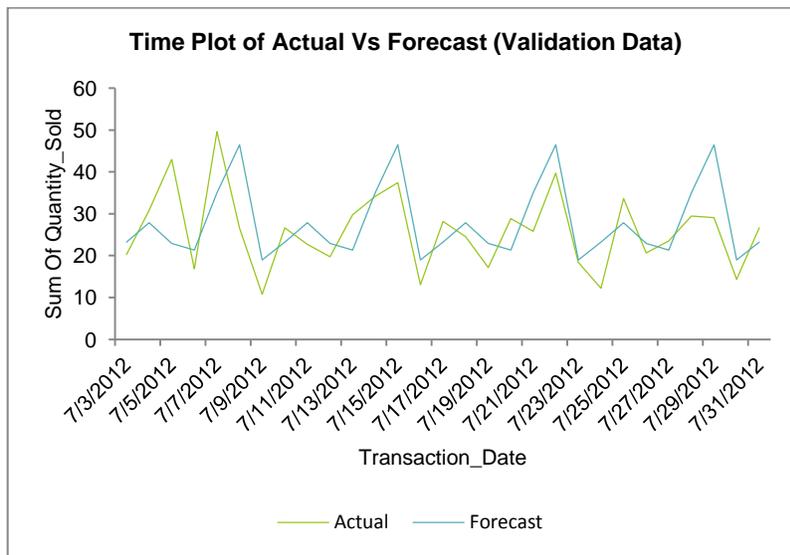
Observing the series for both beans and exotic vegetables, we inferred that there was no observable trend that could affect the forecasting model in the future. Specifically, since we have just one year of data, to infer an annual trend for beans maybe flawed.



## Forecasting Exotic Vegetables

### STEP 1: HOLT WINTER MODEL, NO TREND, PERIOD 7 FOR EXOTIC

Given that the Exotic vegetables had a seasonality of 7 and no observable trend, we decided to take the last 1 year data and forecast daily sales for the month of July.



MAE 4.89  
RMSE 71.28

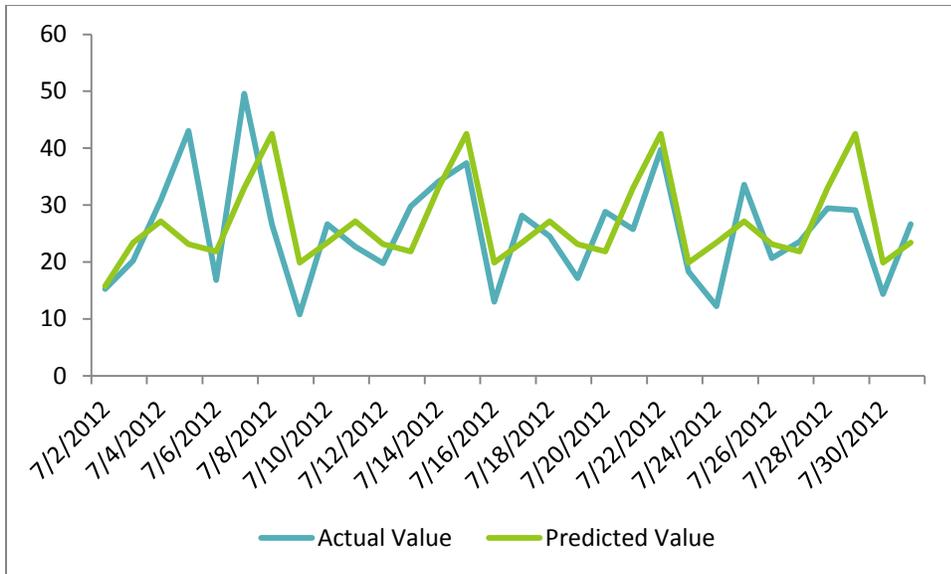
The model predicted with a fairly good accuracy what the peaks look like. Leveraging the cost model that we had earlier, we got an RMSE error cost of 71.28 that comfortably beat the benchmark.

However, we were unsatisfied with the model as such since the peaks were being missed by a fair margin in all the cases.

### STEP 2: LINEAR REGRESSION USING ERRORS AND RELEVANT HOLIDAYS

We took the announced bank holidays and ignored the ones that landed on weekends such that we had a clear idea of which events could potentially be affecting peaks. Now, we ran a regression model such that these relevant holidays were also included plus the residuals from the step 1 model.

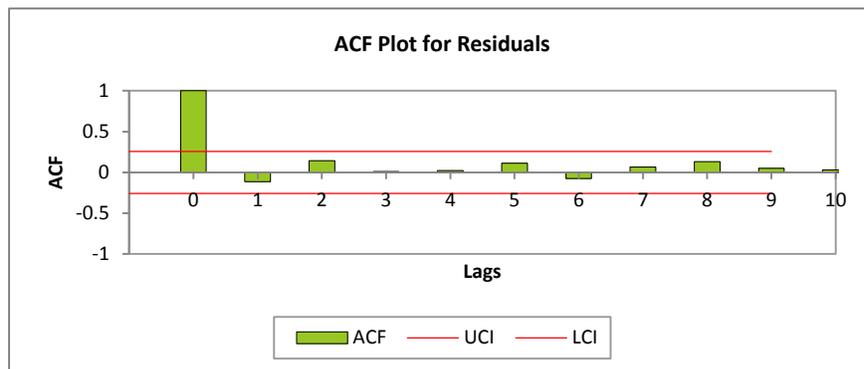
The results were improved significantly accounting for this as shown below:



MAE 2.88  
RMSE 29.85

Now, there were only 2 peaks left where we were completely off. We did a press search to realize that July 4 and July 27 were very **high rainfall days in Mumbai**. Our hypothesis is that due to the heavy rain on July 4, there was a stockout of vegetables on 5 (maybe due to late delivery) and on July 27, the rains did not let shoppers come out in large numbers to visit the store resulting in poor sales.

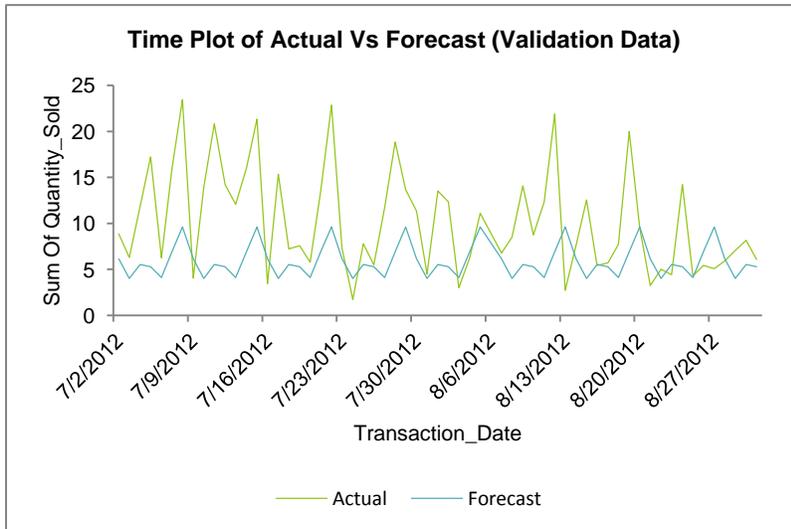
Post the forecast, we observed the residuals to infer that the errors had no observable trend / seasonality implying that it shall be hard to better this.



## Forecasting Beans

### STEP 1: HOLT WINTER MODEL, NO TREND, PERIOD 7 FOR BEAN

Following a similar approach as earlier, we fit a holt-winter model and predicted the July data. In this model, the Holt winter gave a fairly poor result. However, the model consistently under-predicts and is just accurately forecasting data peaks

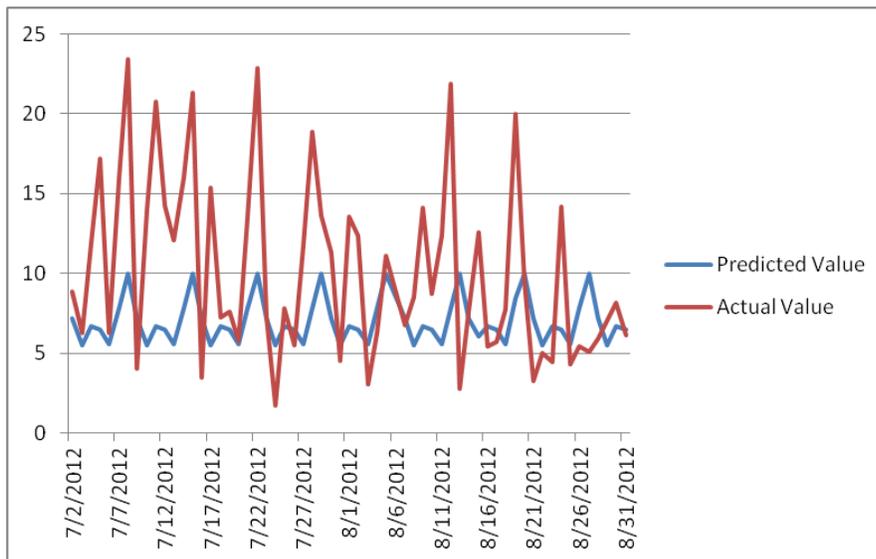


MAE 4.72  
RMSE 51.40

## STEP 2: LINEAR REGRESSION USING ERRORS AND RELEVANT HOLIDAYS

Again, with an approach such as for the exotic above, we fit the errors and relevant holidays to improve the fit. What turned out is an approximately better fit of the actual data with the mean absolute error and the root mean squared error much lower than the benchmark naive forecast.

Although the results obtained from this model are not the best fit for the data, we have chosen the model which gave us the least error after factoring in the cost of over prediction and under prediction. As this model is very rarely over predicting we can compromise on the accuracy is the cost of lost revenues in this term is comparatively lower.



MAE 3.938  
RMSE 5.898

# Key Learning & Observations

---

**Final Model Chosen:** Eventually after multiple rounds of trial and experimentation we consented on going with the two step model of using the Holt-Winter's for the first step followed by running a Linear regression by factoring in the Holt-Winter's forecasts along with holidays not falling on weekends. These holidays which mattered to us were factored into our model in the form of a dummy variable and included in the regression model.

**Possible Alternatives:** Since our historical annual data did not show any clear trend but showed substantial weekly seasonality in the form of peaked sales on weekends followed by relatively higher sales on a Wednesday, we could have run a linear regression capturing seasonality by creating 7 dummy variables.

However this model gave extremely skewed results for the validation data even after running well on the Training data. This drove us towards completely rejecting the model and sticking to Holt-Winter's model without trend.

Apart from the above technical learning we had the following three major insights in terms of the business goal of our project and the interpretation of our given data.

**Day of Week seasonality:** Our historical data clearly showed that sales are highest on Sundays followed by Saturdays and Wednesdays. This seasonality was perfectly replicated week after week and this was also confirmed by plotting an autocorrelation plot which should high positive correlation at lag 7

**Holidays:** Another observation indicated that Middle of the week holidays can result in an increase in the quantity sold in the particular week thus resulting in distortion of the current seasonality. For this we factored the holidays in our regression model by creating a dummy variable

**Rainfall:** Modeling rainfall data can help improve predictive ability of the model as a few days in July and Aug which experienced heavy rainfalls disrupted the regular occurrence of events. This

resulted in under-prediction when the sales were lower than expected as customers were unable to travel due to heavy rains. Secondly it resulted in over-predictions when due to heavy rains the earlier day the supplier was unable to supply the expected quantity.

**Spoilage costs:** High demand volatility translates to high business risk and forecasting models should penalize overstocking to make the most effective use of this forecasting model.

## Exhibits

---

### Exhibit 1: Exotic Vegetables Forecast Data after Step 2 of regression

| Transaction_Date | Predicted Value | Actual Value |
|------------------|-----------------|--------------|
| 7/2/2012         | 15.774119       | 15.246       |
| 7/3/2012         | 23.417223       | 20.232       |
| 7/4/2012         | 27.179446       | 30.858       |
| 7/5/2012         | 23.140269       | 43.008       |
| 7/6/2012         | 21.851828       | 16.812       |
| 7/7/2012         | 33.045894       | 49.572       |
| 7/8/2012         | 42.519984       | 26.598       |
| 7/9/2012         | 19.890726       | 10.764       |
| 7/10/2012        | 23.417223       | 26.658       |
| 7/11/2012        | 27.179446       | 22.668       |
| 7/12/2012        | 23.140269       | 19.752       |
| 7/13/2012        | 21.851828       | 29.778       |
| 7/14/2012        | 33.045894       | 34.152       |
| 7/15/2012        | 42.519984       | 37.41        |
| 7/16/2012        | 19.890726       | 13.014       |
| 7/17/2012        | 23.417223       | 28.14        |
| 7/18/2012        | 27.179446       | 24.54        |
| 7/19/2012        | 23.140269       | 17.172       |
| 7/20/2012        | 21.851828       | 28.836       |
| 7/21/2012        | 33.045894       | 25.782       |
| 7/22/2012        | 42.519984       | 39.714       |
| 7/23/2012        | 19.890726       | 18.342       |
| 7/24/2012        | 23.417223       | 12.216       |
| 7/25/2012        | 27.179446       | 33.606       |
| 7/26/2012        | 23.140269       | 20.646       |

|           |           |        |
|-----------|-----------|--------|
| 7/27/2012 | 21.851828 | 23.568 |
| 7/28/2012 | 33.045894 | 29.466 |
| 7/29/2012 | 42.519984 | 29.112 |
| 7/30/2012 | 19.890726 | 14.37  |
| 7/31/2012 | 23.417223 | 26.67  |

**Exhibit 2: Beans Forecast data after Step 2 of Regression**

| Transaction_Date | Predicted Value | Actual Value |
|------------------|-----------------|--------------|
| 7/2/2012         | 7.2112126       | 8.862        |
| 7/3/2012         | 5.4617926       | 6.282        |
| 7/4/2012         | 6.7087222       | 11.874       |
| 7/5/2012         | 6.492839        | 17.22        |
| 7/6/2012         | 5.5305779       | 6.246        |
| 7/7/2012         | 7.8307532       | 15.942       |
| 7/8/2012         | 10.00824        | 23.442       |
| 7/9/2012         | 7.2112126       | 4.05         |
| 7/10/2012        | 5.4617926       | 13.938       |
| 7/11/2012        | 6.7087222       | 20.796       |
| 7/12/2012        | 6.492839        | 14.22        |
| 7/13/2012        | 5.5305779       | 12.066       |
| 7/14/2012        | 7.8307532       | 16.02        |
| 7/15/2012        | 10.00824        | 21.324       |
| 7/16/2012        | 7.2112126       | 3.444        |
| 7/17/2012        | 5.4617926       | 15.342       |
| 7/18/2012        | 6.7087222       | 7.224        |
| 7/19/2012        | 6.492839        | 7.572        |
| 7/20/2012        | 5.5305779       | 5.76         |
| 7/21/2012        | 7.8307532       | 13.47        |
| 7/22/2012        | 10.00824        | 22.866       |
| 7/23/2012        | 7.2112126       | 7.284        |
| 7/24/2012        | 5.4617926       | 1.722        |
| 7/25/2012        | 6.7087222       | 7.776        |
| 7/26/2012        | 6.492839        | 5.49         |
| 7/27/2012        | 5.5305779       | 11.712       |
| 7/28/2012        | 7.8307532       | 18.84        |
| 7/29/2012        | 10.00824        | 13.65        |
| 7/30/2012        | 7.2112126       | 11.34        |

|           |           |        |
|-----------|-----------|--------|
| 7/31/2012 | 5.4617926 | 4.482  |
| 8/1/2012  | 6.7087222 | 13.524 |
| 8/2/2012  | 6.492839  | 12.348 |
| 8/3/2012  | 5.5305779 | 3.018  |
| 8/4/2012  | 7.8307532 | 6.276  |
| 8/5/2012  | 10.00824  | 11.106 |
| 8/7/2012  | 7.2112126 | 6.768  |
| 8/8/2012  | 5.4617926 | 8.496  |
| 8/9/2012  | 6.7087222 | 14.076 |
| 8/10/2012 | 6.492839  | 8.736  |
| 8/11/2012 | 5.5305779 | 12.33  |
| 8/12/2012 | 7.8307532 | 21.894 |
| 8/13/2012 | 10.00824  | 2.736  |
| 8/14/2012 | 7.2112126 | 7.62   |
| 8/15/2012 | 6.0603796 | 12.54  |
| 8/16/2012 | 6.7087222 | 5.442  |
| 8/17/2012 | 6.492839  | 5.694  |
| 8/18/2012 | 5.5305779 | 7.728  |
| 8/19/2012 | 8.4293401 | 19.974 |
| 8/20/2012 | 10.00824  | 9.504  |
| 8/21/2012 | 7.2112126 | 3.264  |
| 8/22/2012 | 5.4617926 | 5.022  |
| 8/23/2012 | 6.7087222 | 4.416  |
| 8/24/2012 | 6.492839  | 14.202 |
| 8/25/2012 | 5.5305779 | 4.296  |
| 8/26/2012 | 7.8307532 | 5.424  |
| 8/27/2012 | 10.00824  | 5.07   |
| 8/28/2012 | 7.2112126 | 5.916  |
| 8/29/2012 | 5.4617926 | 7.068  |
| 8/30/2012 | 6.7087222 | 8.154  |
| 8/31/2012 | 6.492839  | 6.096  |

**Exhibit 3: Cost-Residual Linkage and calculation for Exotic Vegetables**

| Transaction_Date | Residual   | Cost      |
|------------------|------------|-----------|
| 7/2/2012         | -0.5281194 | 2.725099  |
| 7/3/2012         | -3.1852233 | 0.6727401 |
| 7/4/2012         | 3.6785544  | 26.680094 |
| 7/5/2012         | 19.867731  | 115.07198 |
| 7/6/2012         | -5.0398276 | 0.5118288 |
| 7/7/2012         | 16.526106  | 65.792325 |
| 7/8/2012         | -15.921984 | 180.46591 |
| 7/9/2012         | -9.1267262 | 0.0999327 |
| 7/10/2012        | 3.2407767  | 71.846091 |
| 7/11/2012        | -4.5114456 | 198.45139 |
| 7/12/2012        | -3.3882685 | 59.709017 |
| 7/13/2012        | 7.9261724  | 42.711742 |
| 7/14/2012        | 1.1061064  | 67.063764 |
| 7/15/2012        | -5.109984  | 128.04642 |
| 7/16/2012        | -6.8767262 | 0.1419189 |
| 7/17/2012        | 4.7227767  | 97.618497 |
| 7/18/2012        | -2.6394456 | 0.2655112 |
| 7/19/2012        | -5.9682685 | 1.1645884 |
| 7/20/2012        | 6.9841724  | 0.0526345 |
| 7/21/2012        | -7.2638936 | 31.801105 |
| 7/22/2012        | -2.805984  | 165.32199 |
| 7/23/2012        | -1.5487262 | 0.005298  |
| 7/24/2012        | -11.201223 | 0.1398605 |
| 7/25/2012        | 6.4265544  | 1.1390818 |
| 7/26/2012        | -2.4942685 | 0.0100569 |
| 7/27/2012        | 1.7161724  | 38.209979 |
| 7/28/2012        | -3.5798936 | 121.20352 |
| 7/29/2012        | -13.407984 | 13.262416 |
| 7/30/2012        | -5.5207262 | 17.046885 |
| 7/31/2012        | 3.2527767  | 0.0095999 |

**Exhibit 4: Cost-Residual Linkage and calculation for Beans**

| Transaction_Date | Residual | Cost     |
|------------------|----------|----------|
| 7/2/2012         | 1.65079  | 1.650787 |
| 7/3/2012         | 0.82021  | 0.820207 |
| 7/4/2012         | 5.16528  | 5.165278 |
| 7/5/2012         | 10.7272  | 10.72716 |
| 7/6/2012         | 0.71542  | 0.715422 |
| 7/7/2012         | 8.11125  | 8.111247 |
| 7/8/2012         | 13.4338  | 13.43376 |
| 7/9/2012         | -3.16121 | -0.31612 |
| 7/10/2012        | 8.47621  | 8.476207 |
| 7/11/2012        | 14.0873  | 14.08728 |
| 7/12/2012        | 7.72716  | 7.727161 |
| 7/13/2012        | 6.53542  | 6.535422 |
| 7/14/2012        | 8.18925  | 8.189247 |
| 7/15/2012        | 11.3158  | 11.31576 |
| 7/16/2012        | -3.76721 | -0.37672 |
| 7/17/2012        | 9.88021  | 9.880207 |
| 7/18/2012        | 0.51528  | 0.515278 |
| 7/19/2012        | 1.07916  | 1.079161 |
| 7/20/2012        | 0.22942  | 0.229422 |
| 7/21/2012        | 5.63925  | 5.639247 |
| 7/22/2012        | 12.8578  | 12.85776 |
| 7/23/2012        | 0.07279  | 0.072787 |
| 7/24/2012        | -3.73979 | -0.37398 |
| 7/25/2012        | 1.06728  | 1.067278 |
| 7/26/2012        | -1.00284 | -0.10028 |
| 7/27/2012        | 6.18142  | 6.181422 |
| 7/28/2012        | 11.0092  | 11.00925 |
| 7/29/2012        | 3.64176  | 3.64176  |
| 7/30/2012        | 4.12879  | 4.128787 |
| 7/31/2012        | -0.97979 | -0.09798 |
| 8/1/2012         | 6.81528  | 6.815278 |
| 8/2/2012         | 5.85516  | 5.855161 |
| 8/3/2012         | -2.51258 | -0.25126 |
| 8/4/2012         | -1.55475 | -0.15548 |
| 8/5/2012         | 1.09776  | 1.09776  |
| 8/7/2012         | -0.44321 | -0.04432 |
| 8/8/2012         | 3.03421  | 3.034207 |
| 8/9/2012         | 7.36728  | 7.367278 |

|           |          |          |
|-----------|----------|----------|
| 8/10/2012 | 2.24316  | 2.243161 |
| 8/11/2012 | 6.79942  | 6.799422 |
| 8/12/2012 | 14.0632  | 14.06325 |
| 8/13/2012 | -7.27224 | -0.72722 |
| 8/14/2012 | 0.40879  | 0.408787 |
| 8/15/2012 | 6.47962  | 6.47962  |
| 8/16/2012 | -1.26672 | -0.12667 |
| 8/17/2012 | -0.79884 | -0.07988 |
| 8/18/2012 | 2.19742  | 2.197422 |
| 8/19/2012 | 11.5447  | 11.54466 |
| 8/20/2012 | -0.50424 | -0.05042 |
| 8/21/2012 | -3.94721 | -0.39472 |
| 8/22/2012 | -0.43979 | -0.04398 |
| 8/23/2012 | -2.29272 | -0.22927 |
| 8/24/2012 | 7.70916  | 7.709161 |
| 8/25/2012 | -1.23458 | -0.12346 |
| 8/26/2012 | -2.40675 | -0.24068 |
| 8/27/2012 | -4.93824 | -0.49382 |
| 8/28/2012 | -1.29521 | -0.12952 |
| 8/29/2012 | 1.60621  | 1.606207 |
| 8/30/2012 | 1.44528  | 1.445278 |
| 8/31/2012 | -0.39684 | -0.03968 |