MANDI PRICE PREDICTION

BY

SUDHANSU TAPARIA

CONTENTS

PROJECT PURPOSE

The demand for e-market in agriculture is increasing YoY. With it, there is considerable interest in automation of trading and farming. In the developing countries like India, there are no standardized rules for prices/rates for which traders are always at a gain than farmers. For the benefit of farmers, government institutions, and other policy makers we attempt to build a system by taking various factors like rainfall, past mandi pricing, the cost of cultivation to predict the mandi price. A strict and stringent reform policies can be implemented by the policy makers which benefits the farmers in the country.

ABOUT THE DATASET

The primary sources of dataset which were used in our analysis comprised of the

- Mandi Prices Dataset
- Monthly Rainfall Dataset
- Cost of Cultivation Dataset

Mandi Prices Dataset

Mandi Prices Dataset consists of Maximum, Minimum and Mode of Prices at which crops were sold in the market. It also consists of attributes such as Group of the crop which suggest if the crop is a type of cereal or pulses etc., the arrival quantity of the crop in the market and its unit, the unit of the selling price, state name etc.

Monthly Rainfall Dataset

Monthly Rainfall Dataset consists of the average rainfall per month per year for each District. It consists of columns Year, Individual Months, State, District and the Annual Total Rainfall.

Cost of Cultivation Dataset

Cost of Cultivation Dataset consists of various costs involved in cultivation of crops. It consists of columns such as Labor Cost and Labor hours, Machinery Cost and Machinery Hours, Animals Owned Hours and Cost, Total Capital Cost, Insecticide Cost, Manure Cost, Fertilizer Cost Seed, Quantity and Cost etc. The dataset is at a year level for each district and each crop.

ATTRIBUTES

The Mandi Price Dataset has got 33,774,767 records in total with 20 attributes. The Mandi Price Dataset is for data from year 2000-2010 but for year 2000-2003 the data is not available in the sheet. Hence we are only considering the data from 2004-2010

The Monthly Rainfall Dataset has got 707 records in total with 16 attributes. The Monthly Rainfall data is also available from 2004-2010

The Cost of Cultivation Dataset has got 2,42,446 records in total with 78 attributes. The data is available for years from 2004-2012. But since we do not have the mandi prices dataset and the monthly rainfall dataset for 2011-2012 we will be only considering the period of 2004-2010 for our analysis.

Mandi Price Dataset	Monthly Rainfall Dataset	Cost of Cultivation Dataset
Date	State	State
Market	District	Crop
Arrivals	Year	Year
Unit_of_Arr	January	tehsilcultivator
Variety	February	Pps
Min_Prices	March	season
Max_Prices	April	mainprd_qtl
Mod_Prices	May	croparea_ha

Unit_Price	June	mainprd_rs
State	July	byprd_rs
Month	August	famlab_hrs
Year	September	famlab_rs
Dt	October	atchdlab_hrs
missing1	November	atchdlab_rs
c_miss_missing1	December	casuallab_hrs
Group	Annual Total	casuallab_rs
Commodity		hrdanimllab_hrs
Mth		hrdanimllab_rs
n		ownanimllab_hrs
nn		ownanimllab_rs
		hrdmchn_hrs
		hrdmchn_rs
		ownmchn_hrs
		ownmchn_rs
		seed_kg
		seed_rs
		fertn_kg
		fertn_rs
		fertp_kg
		fertp_rs
		fertk_kg
		fertk_rs
		fertoth_kg

		fertoth_rs
		ferttotal_kg
		ferttotal_rs
		Zone
		tehsil_code_d
		Sizegroup
		manure_qtl
		manure_rs
		insecticide_rs
		ownirrimchn_hrs
		ownirrimchn_rs
		hrdirrimchn_hrs
		hrdirrimchn_rs
		misc_rs
		landrevenue_rs
		rpll_rs
		imputedrent_rs
		totaldepre_rs
		totalcapital_rs
		ss_groupno
		area_sel_cr_vil_ha
		nvillages_tehsil
		ngrowers_cluster
		cluster_weight

		cropareainzone_ha
		cropprodinzone_qtl
		Ntehsilsinzone
		n_samp_teh_zone
		area_sel_cr_zn_ha
		Minrent
		Maxrent
		samp_zo_state
		samp_cl_state
		Tenure
		Variety
		canalandothirri_rs
		state_code
		crop_code
		zone_cb
		district_cb
		tehsil_code_cb
		tehsil_cb
		state_11
		district_code
		district_11

CHALLENGES FACED

To build an advanced analytical model to perform Price Prediction Analysis and Clustering Analysis for the market and crop groups we wanted to relate the data and get rid of impurities in the dataset.

Major Challenges that the team faced were:

- The Mandi Price Dataset was not having pricing details for year from 2000-2003
- The Mandi Price Dataset had a lot of junk values (e.g. A Price field consists of values like N.R, blanks values and null values)
- The Price for the commodity in the dataset was in Rupee/Quintal but the arrival quantity was in Tonne.
- The Monthly Rainfall dataset was having columns as Jan, Feb, March month names which needed to be unpivoted.
- The Cost of Cultivation Dataset had district codes instead of actual names which had to be mapped to the Mandi Price Dataset and Monthly Rainfall Dataset.

SOLUTION TO THE CHALLENGES FACED

- **The Mandi Price Dataset was not having pricing details for year from 2000-2003**

  We had to let go of the rows for the year from 2000-2003 as there was no data which was available for those days. We only utilized the data from 2004-2010

- **The Mandi Price Dataset had a lot of junk values (e.g. A Price field consists of values like N.R, blanks values and null values)**

  We replaced the N.R. values with 0 instead to make it possible to do aggregations. The blank values were replaced and set to null values. The null values were then handled during data pre-processing and were imputed which is described in the section on Exploratory Data Analysis and Data Pre-Processing.

- **The Price for the commodity in the dataset was in Rupee/Quintal but the arrival quantity was in Tonne.**

  We multiplied the Price of the commodity by 10 to bring it Rupee/Tonne which would be our unit in this study.

- **The Monthly Rainfall dataset was having columns as Jan, Feb, March month names which needed to be unpivoted.**

  The monthly Rainfall Dataset was unpivoted to have the Months on rows instead of columns and can analyze.

- **The Cost of Cultivation Dataset had district codes instead of actual names which had to be mapped to the Mandi Price Dataset and Monthly Rainfall Dataset.**

  We created a bridge table consisting of District Codes and their Names and then connected the Cost of Cultivation Dataset to Mandi Price Dataset and Rainfall Dataset.

EXPLORATORY DATA ANALYSIS

In this section of the document we will focus on the exploratory data analysis of the Mandi Prices Dataset, Monthly Rainfall Dataset and the Cost of Cultivation dataset. We shall also look at the preprocessing/ cleaning performed on the same.

IMPUTATION FOR MISSING VALUES

The below mentioned numeric attributes contain Null values which are impute using the following approach.

For which ever district the rainfall data was NULL, the same was replaced by the average of the rainfall in that State in that month if the same amount of rainfall will happen in that area too.

For which ever district the Cost of Cultivation dataset was NULL, the same was replaced by the average of that price for that crop group in that state in the same month of the year.

VARIABLES SELECTED FOR MODELLING

The following table shows the complete list of variables:

Variable Name	Description	Distinct Count	Mean
PRICE	Price of the crop	63083	181.51
Market	Market Name	820	NA
State	State Name	4	NA
Year	Year	7	NA
Month	Month of Year	12	NA
Group	Crop Group	13	NA
QUANTITY	Quantity of Arrival	57716	171.35
RAINFALL	Amount of Rainfall	1739	71.99
MainProduct_Tonne	Main Product in tonne	788	841.12
CropArea_Hectare	Crop Area in Hectare	749	1.026
FarmLabour_Hours	Farm Labor Hours	791	196.32
FarmLabour_Cost	Farm Labor Cost	792	2185.8
AnimalOwned_Hours	Animal Owned Hour	723	27.99
AnimalOwned_Cost	Animal Owned Cost	772	1105.58

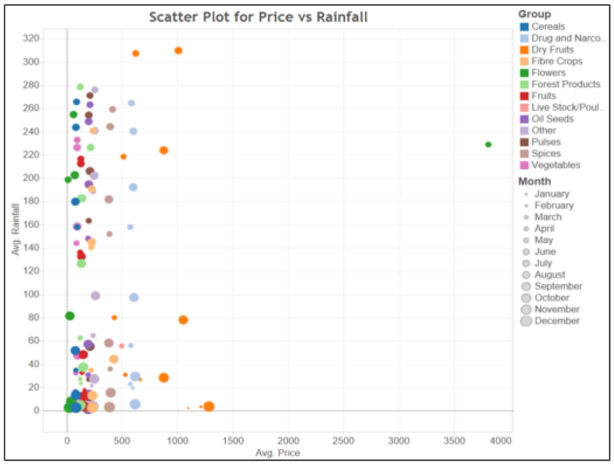
Machinery_Hours	Machinery Hours	469	6.2
Machinery_Cost	Machinery Cost	481	816.09
SeedQuantity_Kg	Seed Quantity in Kg	693	157.95
Seed_Cost	Seed Cost	763	2698.24
FertilizerQuantity_Kg	Fertilizer Quantity in Kg	764	93.79
Fertilizer_Cost	Fertilizer Cost	785	996.49
Manure_Tonne	Manure Quantity in Tonne	530	101.53
Manure_Cost	Manure Cost	534	3245.53
Insectiside_Cost	Insecticide Cost	568	660.51
TotalCapital_Cost	Total Capital Cost	790	101638.55

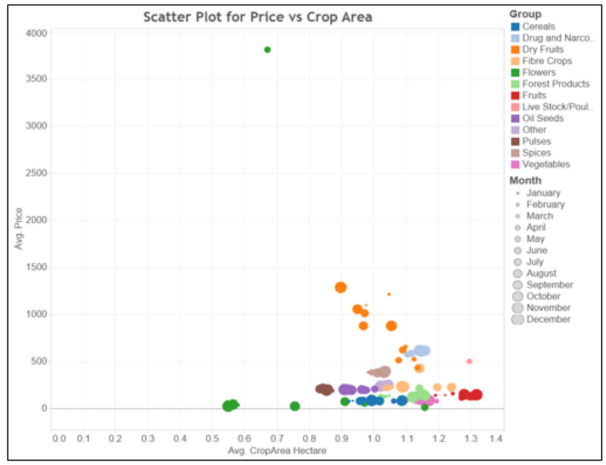
SUMMARY STATISTICS OF VARIABLES

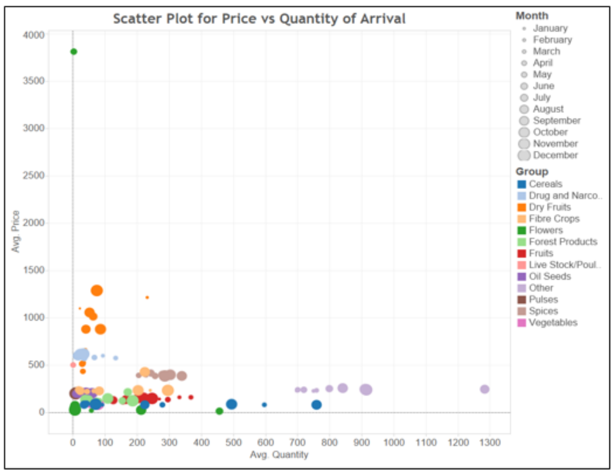
```
                 Market                State             Year           Month                Group            PRICE
 Aurangabad          :  768   Gujarat    :19914   Min.   :2004   February: 9909   Cereals   :23711   Min.   :     0.10
 Shimoga             :  583   Karnataka  :34074   1st Qu.:2005   January : 9804   Vegetables:16298   1st Qu.:    80.29
 Mysore (Bandipalya):  566   Maharashtra:33490   Median :2007   March   : 9384   Oil Seeds :15255   Median :   141.44
 Bangalore           :  552   Punjab     :15292   Mean   :2007   April   : 9307   Pulses    :15039   Mean   :   181.51
 Tumkur              :  547                       3rd Qu.:2008   May     : 8873   Spices    : 8193   3rd Qu.:   222.28
 Davangere           :  540                       Max.   :2010   October : 8738   Fruits    : 7920   Max.   :116180.31
 (Other)             :99214                                      (Other) :46755   (Other)   :16354
    QUANTITY              RAINFALL         MainProduct_Tonne CropArea_Hectare FarmLabour_Hours FarmLabour_Cost AnimalOwned_Hours
 Min.   :     0.01   Min.   :   0.00   Min.   :    3.5   Min.   :0.0500   Min.   :   0.0   Min.   :    0   Min.   :  0.00
 1st Qu.:     1.91   1st Qu.:   0.38   1st Qu.:  645.6   1st Qu.:0.6499   1st Qu.: 131.8   1st Qu.: 1095   1st Qu.: 25.63
 Median :     8.18   Median :  14.12   Median :  841.1   Median :1.0268   Median : 196.3   Median : 2186   Median : 28.00
 Mean   :   171.36   Mean   :  72.00   Mean   :  841.1   Mean   :1.0268   Mean   : 196.3   Mean   : 2186   Mean   : 28.00
 3rd Qu.:    33.33   3rd Qu.: 122.27   3rd Qu.:  992.0   3rd Qu.:1.2671   3rd Qu.: 263.5   3rd Qu.: 2906   3rd Qu.: 29.22
 Max.   :259213.75   Max.   :2281.50   Max.   :16200.0   Max.   :3.1650   Max.   :1191.6   Max.   :14805   Max.   :248.00

 AnimalOwned_Cost Machinery_Hours  Machinery_Cost    SeedQuantity_Kg     Seed_Cost      FertilizerQuantity_Kg Fertilizer_Cost
 Min.   :   0.0   Min.   : 0.000   Min.   :    0.0   Min.   :   0.00   Min.   :    0   Min.   :  0.00        Min.   :   0.0
 1st Qu.: 975.2   1st Qu.: 1.242   1st Qu.:  104.5   1st Qu.:  27.03   1st Qu.: 1454   1st Qu.: 37.80        1st Qu.: 406.3
 Median :1105.6   Median : 3.441   Median :  269.0   Median :  79.24   Median : 2178   Median : 70.48        Median : 834.1
 Mean   :1105.6   Mean   : 6.201   Mean   :  816.1   Mean   : 157.95   Mean   : 2698   Mean   : 93.80        Mean   : 996.5
 3rd Qu.:1390.5   3rd Qu.: 6.201   3rd Qu.:  816.1   3rd Qu.: 157.95   3rd Qu.: 2698   3rd Qu.: 93.80        3rd Qu.: 996.5
 Max.   :7776.0   Max.   :69.524   Max.   :14957.7   Max.   :8225.00   Max.   :82225   Max.   :822.00        Max.   :7885.4

  Manure_Tonne      Manure_Cost     Insectiside_Cost TotalCapital_Cost
 Min.   :   0.00   Min.   :    0   Min.   :   0.0   Min.   :   600
 1st Qu.:  47.06   1st Qu.: 2406   1st Qu.: 108.1   1st Qu.: 59811
 Median :  96.27   Median : 3068   Median : 479.9   Median : 91339
 Mean   : 101.54   Mean   : 3246   Mean   : 660.5   Mean   :101639
 3rd Qu.: 105.42   3rd Qu.: 3246   3rd Qu.: 660.5   3rd Qu.:101639
 Max.   :1600.00   Max.   :89367   Max.   :5516.9   Max.   :539049
```

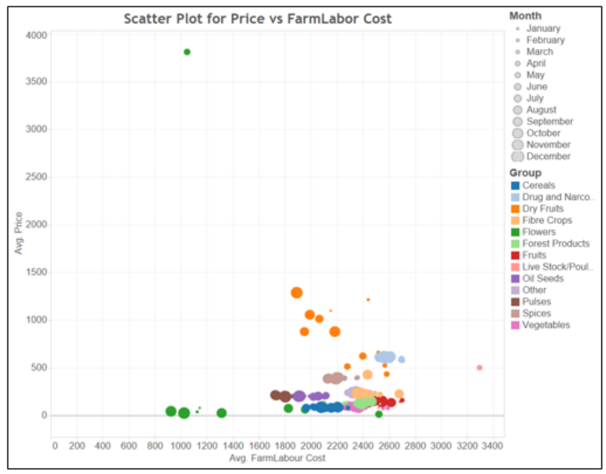
EXPLORATORY DATA ANALYSIS PLOTS

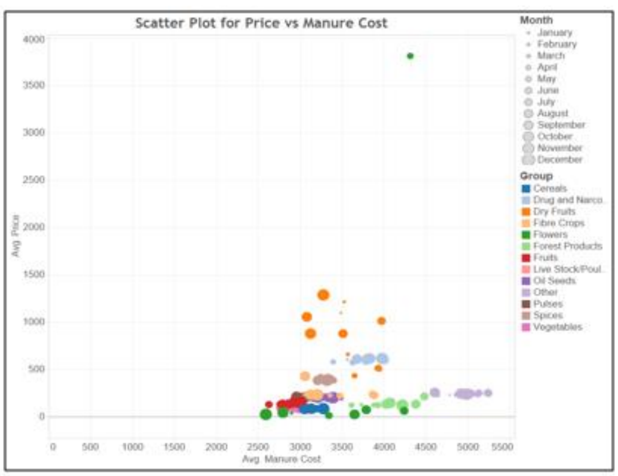
SCATTER PLOTS

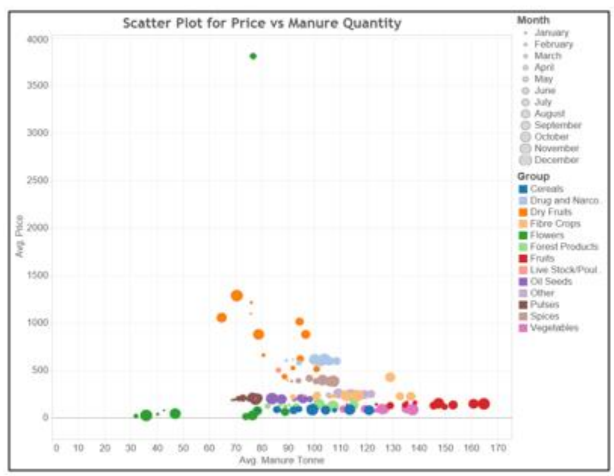


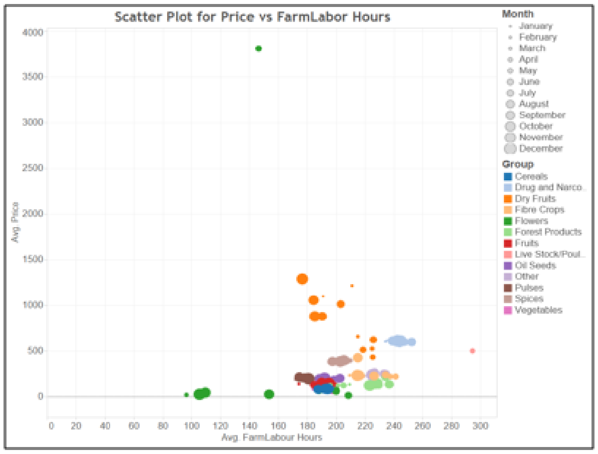


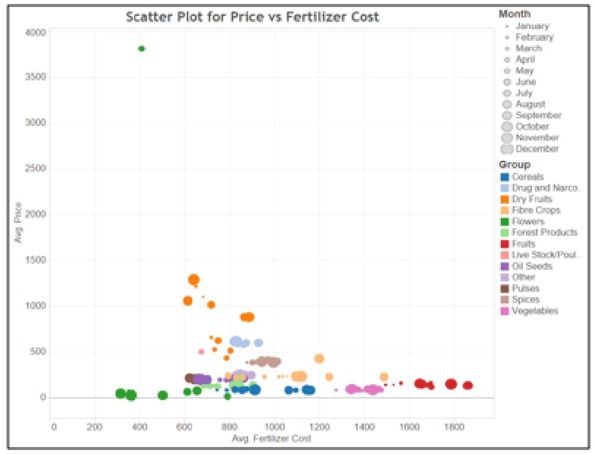


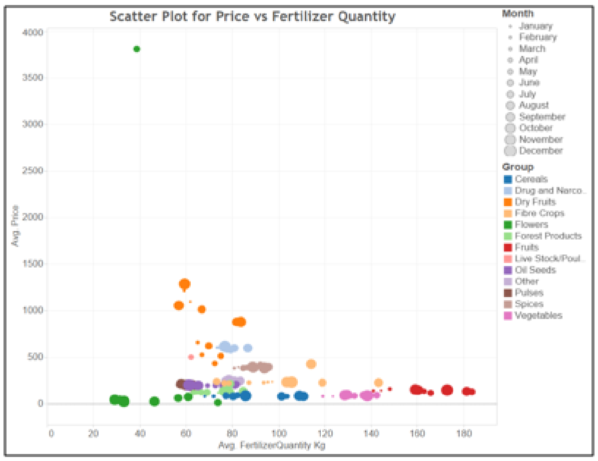


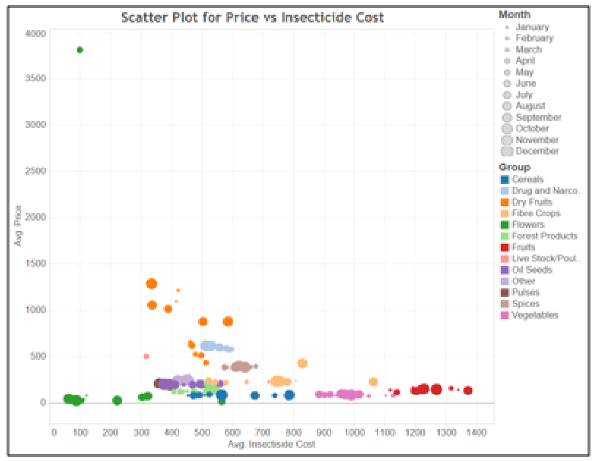


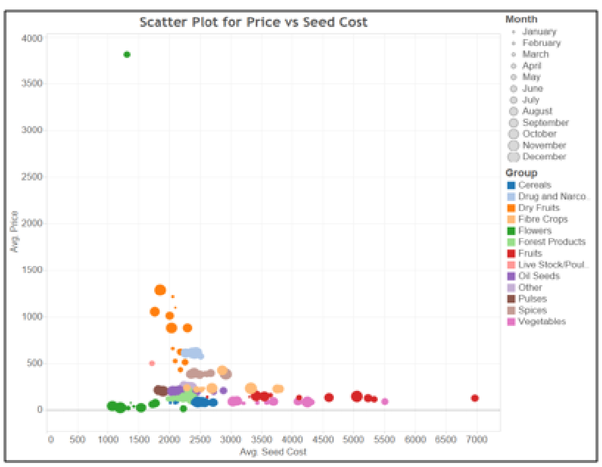


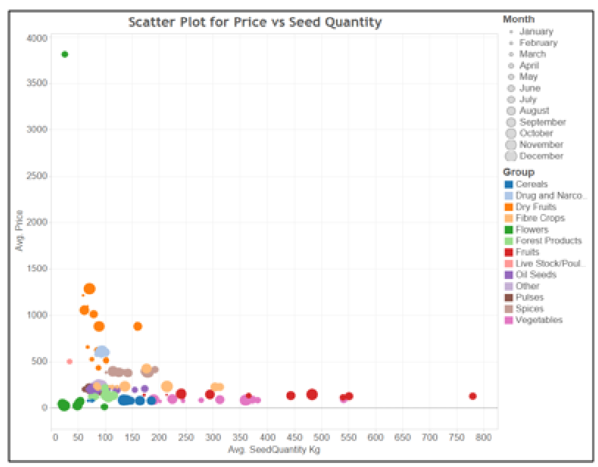

Scatter Plot for Price vs FarmLabor Hours
Avg. Price
Avg. FarmLabour Hours
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables

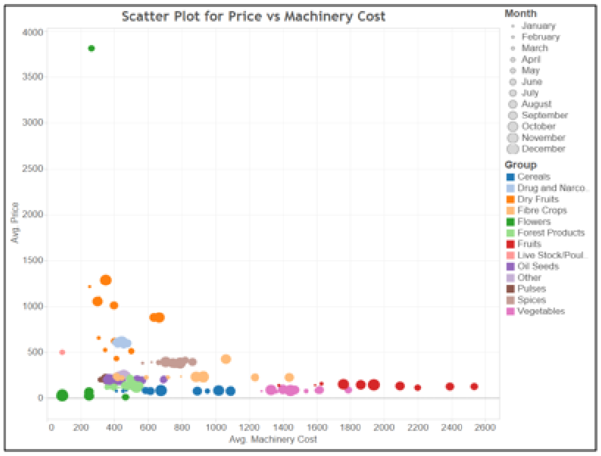

Scatter Plot for Price vs Fertilizer Cost
Avg. Price
Avg. Fertilizer Cost
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables

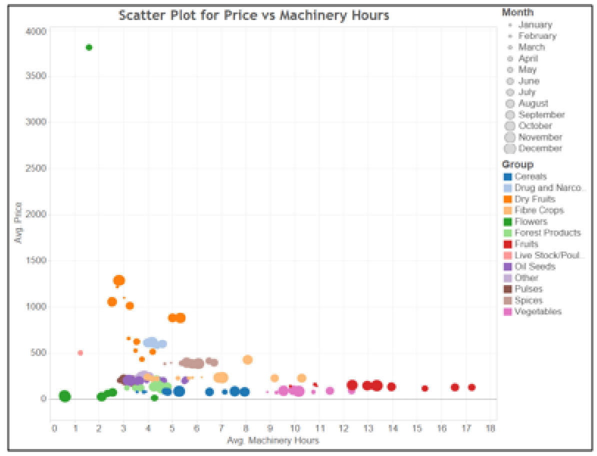

Scatter Plot for Price vs Fertilizer Quantity
Avg. Price
Avg. FertilizerQuantity Kg
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables

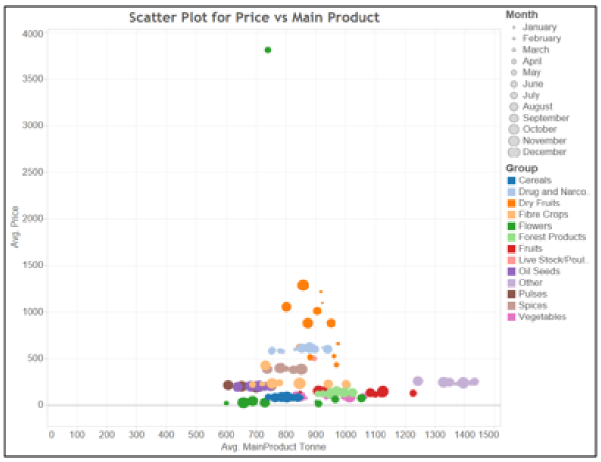

Scatter Plot for Price vs Insecticide Cost
Avg. Price
Avg. Insectiside Cost
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables

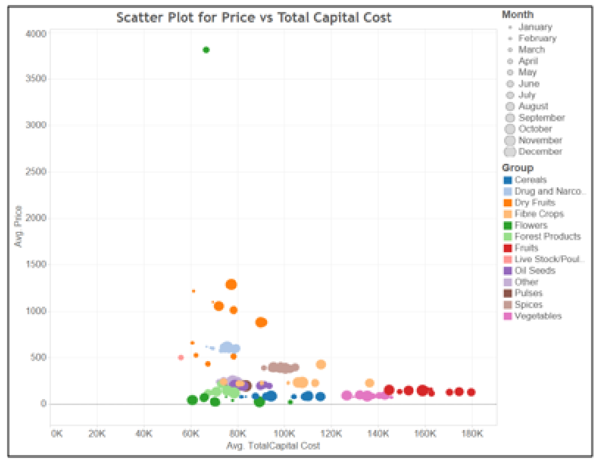

Scatter Plot for Price vs Seed Cost
Avg. Price
Avg. Seed Cost
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables


Scatter Plot for Price vs Seed Quantity
Avg. Price
Avg. SeedQuantity Kg
Month
January
February
March
April
May
June
July
August
September
October
November
December
Group
Cereals
Drug and Narco..
Dry Fruits
Fibre Crops
Flowers
Forest Products
Fruits
Live Stock/Poul..
Oil Seeds
Other
Pulses
Spices
Vegetables

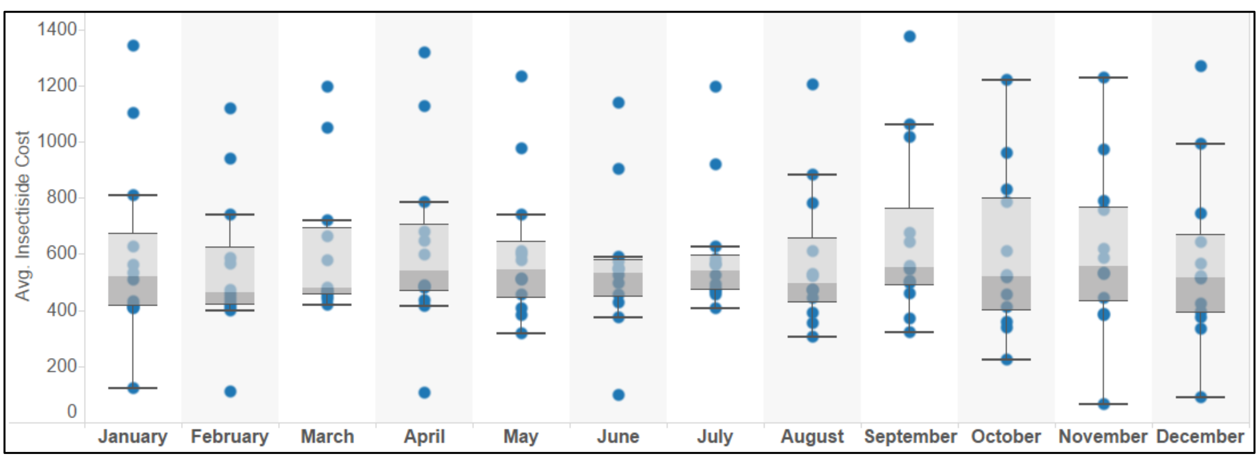


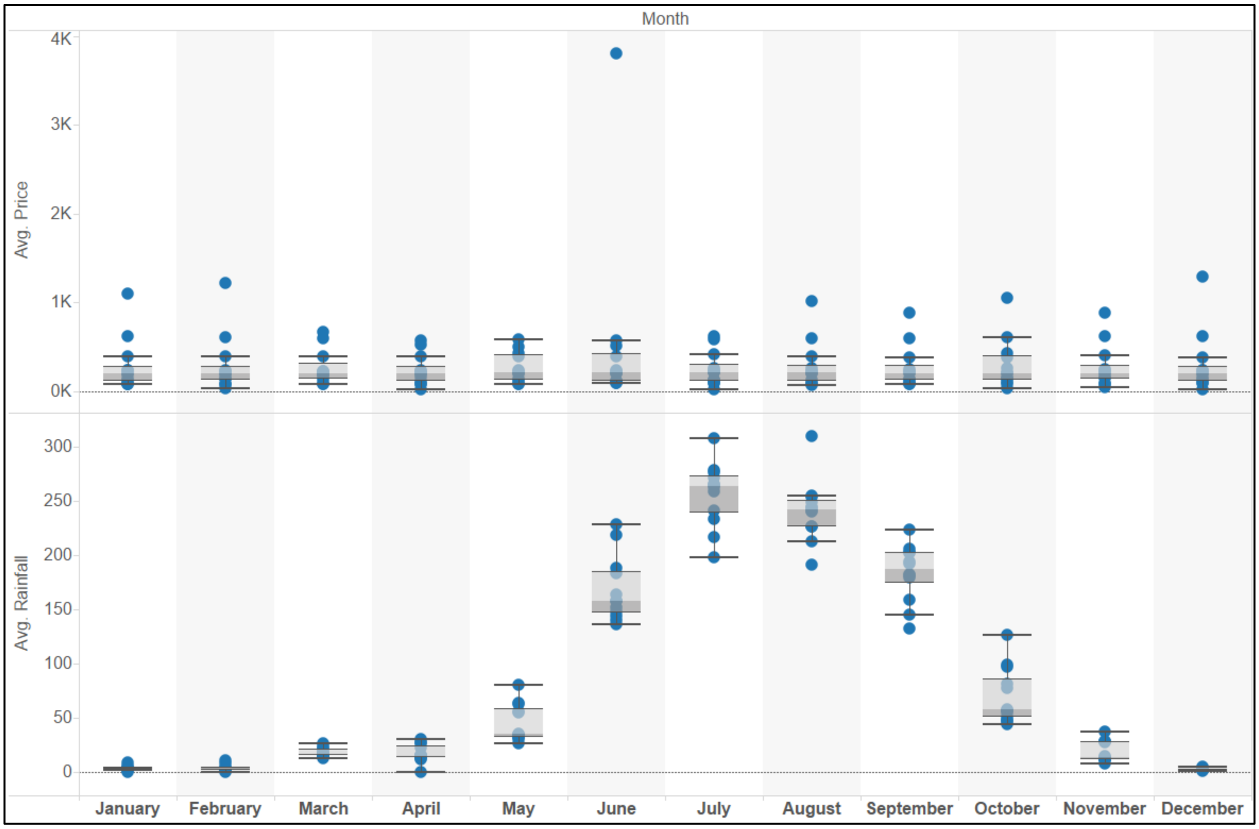


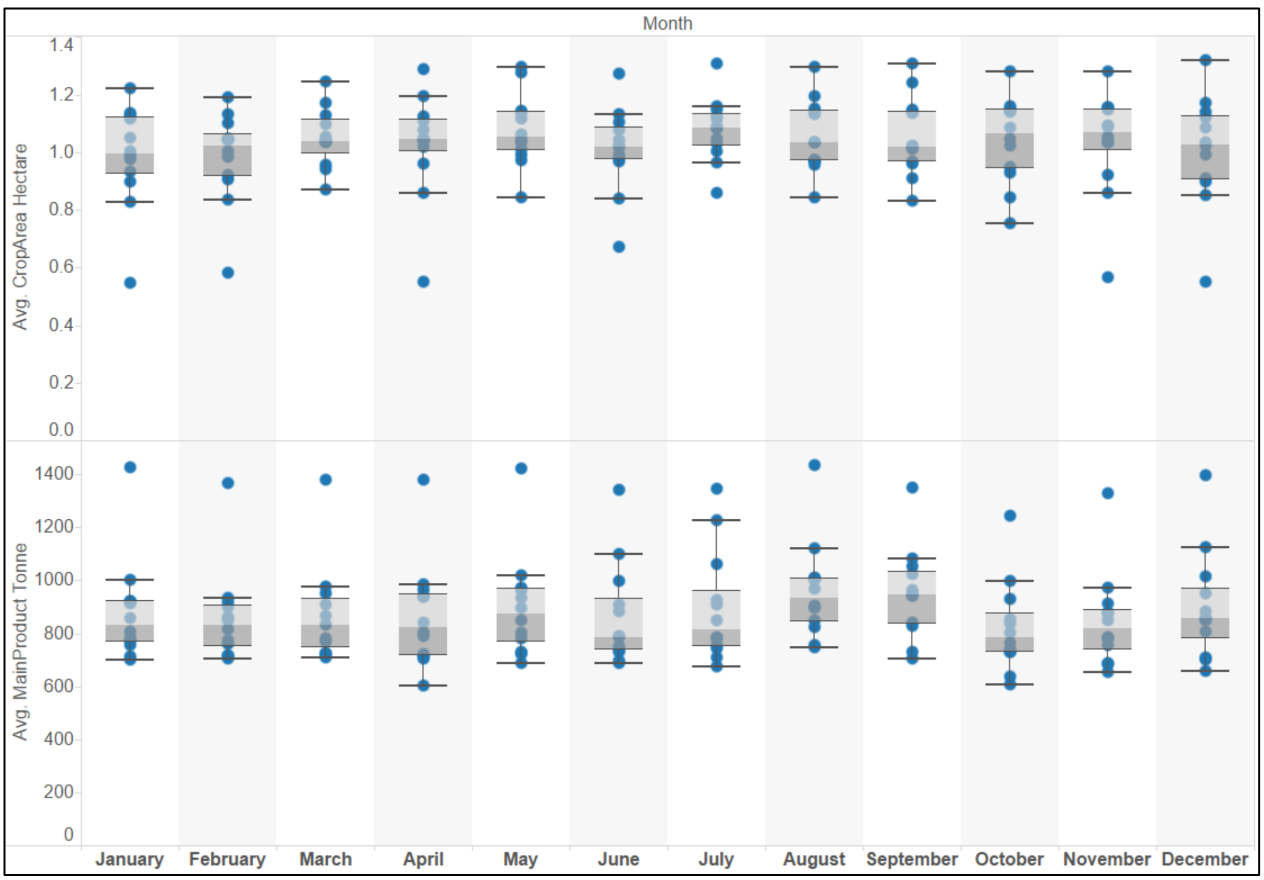


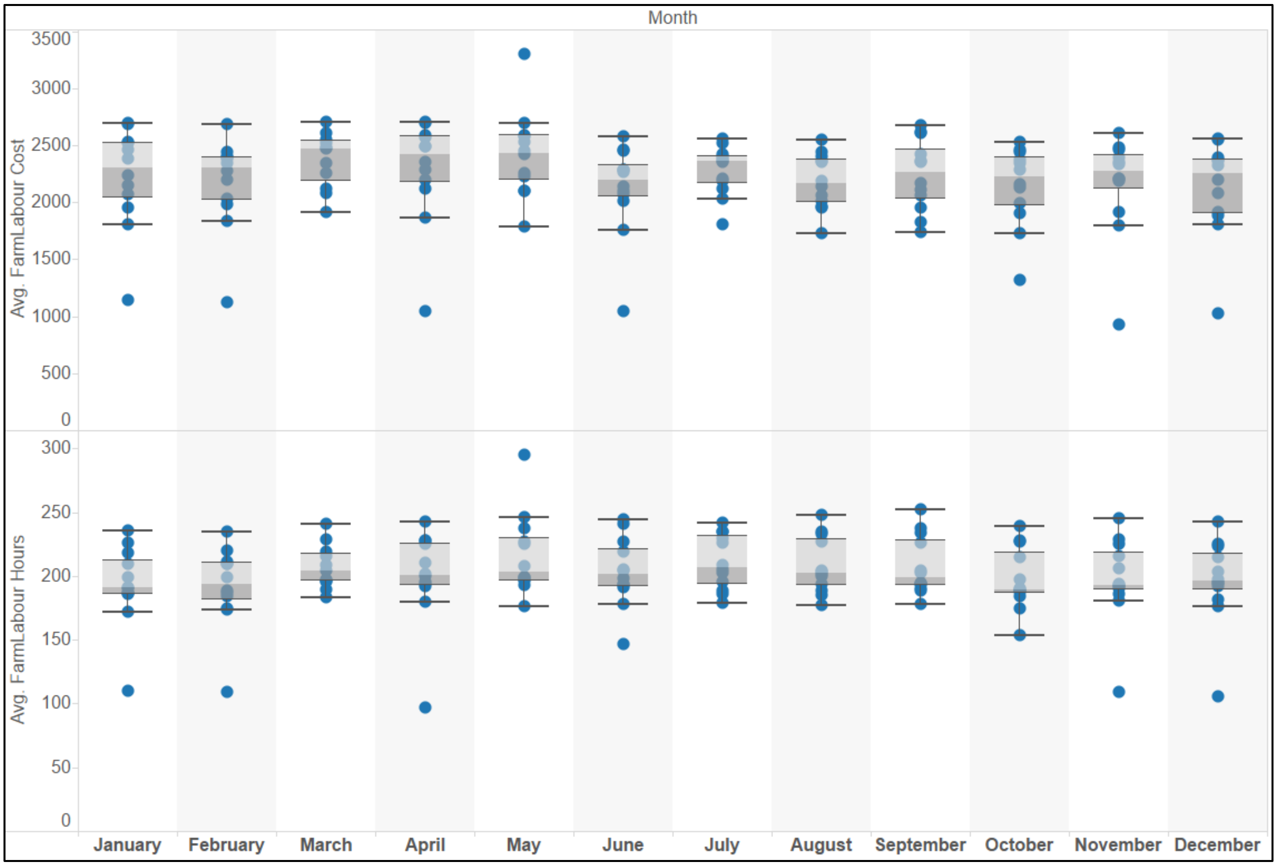


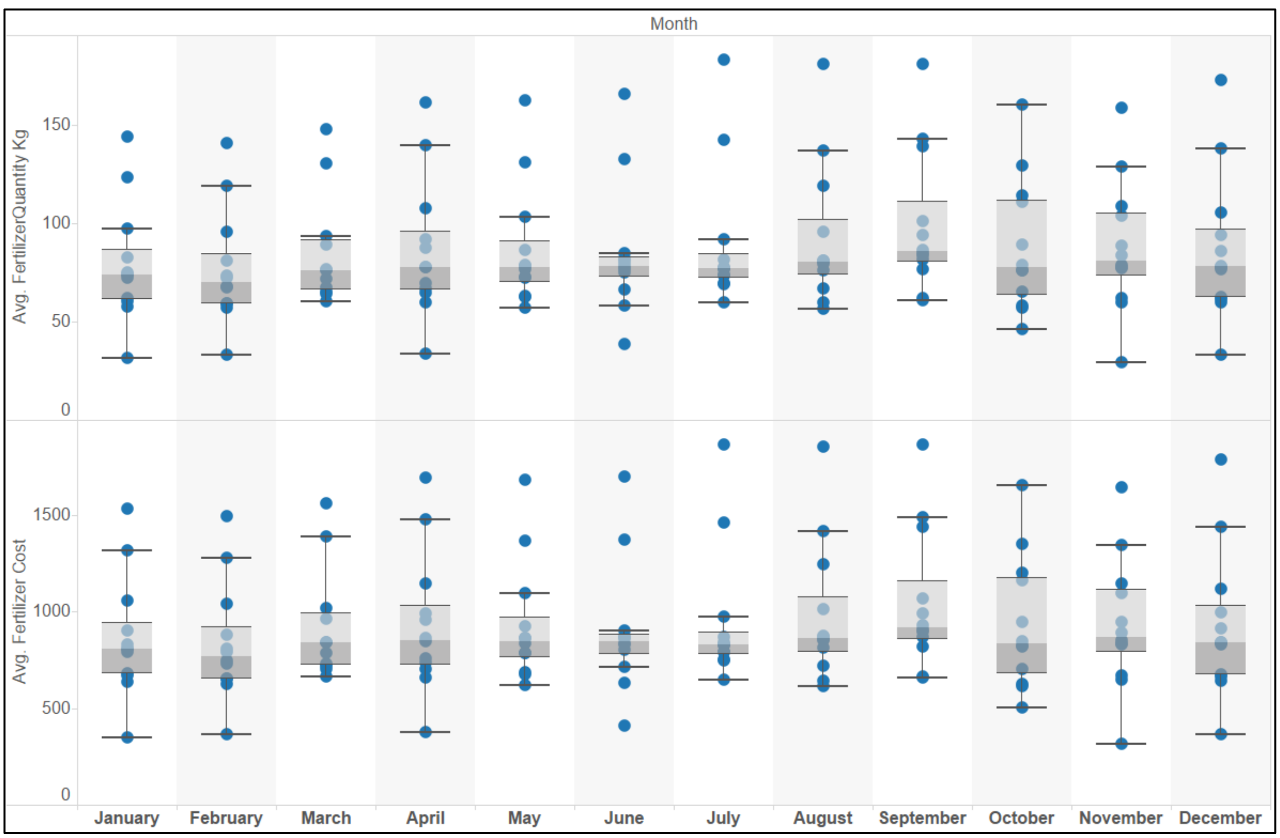
BOX AND WHISKERS PLOT




Month
Avg. Price
4K
3K
2K
1K
0K
Avg. Rainfall
300
250
200
150
100
50
0
January
February
March
April
May
June
July
August
September
October
November
December


Month
Avg. CropArea Hectare
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Avg. MainProduct Tonne
1400
1200
1000
800
600
400
200
0
January
February
March
April
May
June
July
August
September
October
November
December


Month
Avg. FarmLabour Cost
Avg. FarmLabour Hours
January
February
March
April
May
June
July
August
September
October
November
December


Month
Avg. FertilizerQuantity Kg
Avg. Fertilizer Cost
January
February
March
April
May
June
July
August
September
October
November
December

CORRELATION AND COVARIANCE MATRIX

CORRELATION MATRIX

	PRICE	QUANTITY	RAINFALL	MainProduc	CropArea_H	FarmLabour_	FarmLabour_	AnimalOwn	AnimalOwn	Machinery_I	Machinery_(	SeedQuantit	Seed_Cost	FertilizerQu	Fertilizer_Cc	Manure_Tor	Manure_Cos	Insectiside_(	TotalCapital_Cost
PRICE	1																		
QUANTITY	0.0178484	1																	
RAINFALL	0.0105752	-0.010875	1																
MainProduc	-0.019293	0.0245349	0.0162657	1															
CropArea_H	0.0008533	0.0520724	-0.041572	0.2862601	1														
FarmLabour_	0.0456463	0.0192046	0.0370614	0.1562971	0.4426503	1													
FarmLabour_	0.0393874	0.037306	-0.061832	0.1964412	0.7679955	0.6969346	1												
AnimalOwn	0.0315269	-0.010111	0.1058854	-0.019844	-0.245621	0.3441266	-0.197148	1											
AnimalOwn	0.0293084	-0.032055	0.0558358	-0.16075	-0.533295	0.0518404	-0.337422	0.7891735	1										
Machinery_I	-0.04008	0.0318767	-0.07663	0.2748035	0.7639118	0.0264981	0.4894228	-0.62297	-0.719689	1									
Machinery_(	-0.045751	0.0272209	-0.080329	0.3197135	0.7145376	-0.037553	0.423995	-0.637812	-0.705819	0.9817265	1								
SeedQuantit	-0.034399	0.0125508	-0.032836	0.4208903	0.309221	0.0039699	0.0882415	-0.364996	-0.413643	0.5401147	0.639784	1							
Seed_Cost	-0.022181	0.0050514	-0.048616	0.4305718	0.3671284	0.1347399	0.2075952	-0.321078	-0.372628	0.5284082	0.6281042	0.9422673	1						
FertilizerQu	-0.037855	0.0341466	-0.082971	0.3391722	0.8017028	0.1095896	0.5584539	-0.622065	-0.723555	0.9639822	0.9605071	0.5525131	0.5675663	1					
Fertilizer_Cc	-0.036	0.03378	-0.087371	0.3300427	0.8079779	0.1273403	0.5787195	-0.620855	-0.724649	0.9577313	0.9509787	0.5302497	0.5504102	0.9984992	1				
Manure_Tor	-0.026853	0.0334654	-0.038603	0.3986769	0.6383316	0.3450851	0.4157392	-0.147596	-0.413597	0.6499214	0.6229069	0.3187262	0.3569315	0.6731592	0.6746325	1			
Manure_Cos	0.0046422	0.0035165	0.0558308	0.4175393	0.0398953	0.3504094	0.0249076	0.4848952	0.2250442	-0.137469	-0.155618	-0.13158	-0.078135	-0.110918	-0.106738	0.5219753	1		
Insectiside_(	-0.021972	0.0352145	-0.105764	0.1586378	0.8043199	0.1066951	0.6436462	-0.609321	-0.691638	0.9123494	0.8765466	0.3204856	0.3458192	0.9271066	0.9341444	0.6097485	-0.159567	1	
TotalCapital_	-0.049446	0.0187192	-0.089851	0.1901859	0.5958034	-0.131092	0.3570638	-0.71998	-0.698378	0.9034221	0.9201363	0.5133125	0.4812617	0.9092481	0.9057581	0.5568279	-0.237232	0.853514	1

Based on the correlation matrix we can derive a strong relationship between:

- Machinery cost and Machinery hours
- Seed cost and Seed quantity
- Fertilizer quantity with Machinery Hours &Machinery Cost
- Fertilizer cost with Fertilizer Quantity, Machinery Hours & Machinery Cost
- Insecticide cost with Machinery Cost, Fertilizer Quantity, Fertilizer Cost
- Total Capital Cost with Machinery Hours, Machinery Cost, Fertilizer Quantity and Fertilizer Cost

COVARIANCE MATRIX

	PRICE	QUANTITY	RAINFALL	Product_T	Area_Hec	Labour_H	nLabour_C	alOwned_H	alOwned_	chinery_H	achinery_C	dQuantity	Seed_Cost	zerQuantit	rtilizer_Co	nure_Ton	anure_Co	ectiside_C	TotalCapital_Cost
PRICE	165414																		
QUANTITY	11966	2717251																	
RAINFALL	469.847	-1958.31	11933.4																
MainProduct_Tonne	-4893.71	25223.8	1108.2	388977															
CropArea_Hectare	0.13847	34.248	-1.81196	71.2336	0.15919														
FarmLabour_Hours	1338.52	2282.47	291.903	7028.24	12.7338	5198.38													
FarmLabour_Cost	15749.6	60460.1	-6640.82	120454	301.264	49402.8	966611												
AnimalOwned_Hours	188.514	-245.033	170.057	-181.952	-1.4408	364.778	-2849.67	216.149											
AnimalOwned_Cost	6087.77	-26985.9	3115.13	-51202.7	-108.67	1908.9	-169426	5925.55	260831										
Machinery_Hours	-141.931	457.509	-72.8854	1492.26	2.65379	16.6345	4189.59	-79.7452	-3200.26	75.8092									
Machinery_Cost	-25098.9	60525.1	-11836.4	268962	384.552	-3652.09	562282	-12648.4	-486229	11529.7	1819432								
SeedQuantity_Kg	-6111.81	9038.11	-1567.04	114676	53.898	125.043	37900.1	-2344.26	-92288.4	2054.42	377002	190846							
Seed_Cost	-31105.1	28710.7	-18311.8	925918	505.062	33496.2	703733	-16276.2	-656176	15863.4	2921222	1419322	1.2E+07						
FertilizerQuantity_Kg	-1328.97	4858.65	-782.373	18259.4	27.6108	682.036	47393.3	-789.435	-31897.4	724.493	111834	20834.8	168922	7450.89					
Fertilizer_Cost	-12761	48530.4	-8318.34	179399	280.964	8001.82	495887	-7955.28	-322549	7267.64	1117963	201888	1654018	75117.4	759587				
Manure_Tonne	-871.159	4400.3	-336.376	19833.7	20.3156	1984.64	32603.8	-173.09	-16849.2	451.381	67021.2	11106.6	98168.2	4634.93	46900.5	6362.71			
Manure_Cost	4821.57	14803.2	15575.1	665019	40.6499	64518.6	62536.4	18205.4	293510	-3056.62	-536047	-146794	-687999	-24450.2	-237565	106328	6521538		
Insectiside_Cost	-6882.04	44704.1	-8897.8	76195.7	247.146	5924.34	487343	-6898.97	-272032	6117.64	910552	107823	918281	61630.5	626996	37457.1	-313820	593094	
TotalCapital_Cost	-1448669	2222796	-707054	8544539	17124.3	-680864	2.5E+07	-762509	-2.6E+07	566631	8.9E+07	1.6E+07	1.2E+08	5653728	5.7E+07	3199562	-4.4E+07	4.7E+07	5189158225

Based on the covariance matrix we can derive the direction of the linear relationship. Major variables have a positive direction of linear relationship while a few like Rainfall and Quantity, Rainfall and Farm Labor Cost is having a negative direction of linear relationship.

CLUSTERING CHARACTERSTICS OF THE DATASET

CLUSTERING THE GROUP OF CROPS WITH K-MEANS CLUSTERING

We perform the cluster analysis using K-means clustering with K set to 3. We also tried our analysis with K set to other numbers but the Clustering came out to be best with K=3.

We used the below code to perform the analysis.

```
#Load the Datafile
Crop=read.csv("D:/ISB/Capstone/Capstone_PredictionData/CapstonePrediction_Data.csv")
#Copy the  Crop dataset into another dataset
Crop.f=Crop
#Set the categorical variables to null
Crop.f$Market<-NULL
Crop.f$Year<-NULL
Crop.f$Group<-NULL
Crop.f$State<-NULL
Crop.f$Month<-NULL
#Scale the data to normalize
Crop.stand = scale(Crop.f[-1])
#Perform K means clustering
results<-kmeans(Crop.stand,3)
attributes(results)
#Table of Cluster
table(Crop$Group,results$cluster)
```

And below is the cluster table.

	1	2	3
Cereals	18845	2381	2485
Drug and Narcotics	1821	0	529
Dry Fruits	219	2	69
Fibre Crops	4658	666	654
Flowers	33	0	4
Forest Products	1468	4	458
Fruits	4436	2957	527
Live Stock/Poultry/Fisheries	0	0	1
Oil Seeds	13014	411	1830
Other	4263	25	1480
Pulses	13743	160	1136
Spices	6624	673	896
Vegetables	10791	4233	1274

Overall accuracy of the clustering is around 77%.

OUTPUT INTERPRETATION

The K means cluster analysis with K set to 3 does a decent job in most of the cases apart from Vegetables and Cereals where the clusters overlap each other, but clustering in terms of dry fruits, Flowers, Pulses etc. did a good job.

PRICE PREDICTION USING MULTIPLE LINEAR REGRESSION

LOADING THE CSV DATA FILE INTO R STUDIO

```
#Load the Datafile
Crop=read.csv("D:/ISB/Capstone/Capstone_PredictionData/CapstonePrediction_Data.csv")
```

HANDLING CATEGORICAL VARIABLES USING DUMMY VARIABLE

We have a few categorical variables in our dataset like Group which needs to be converted to dummy variables before they can be used in Multiple Linear Regression.

```
new <- dummy.code(Crop$Group)
new.sat <- data.frame(new,Crop)
round(cor(Crop,use="pairwise"),2)
```

LOG TRANSFORMATION OF VARIABLES

The variables in our study are quite varied in terms of their unit and hence they require transformation to bring them to the same scale. We tried both square root transformation and log transformation but ultimately used log transformation as it provided us better results.

We have added 1 to the value because some of the data is 0 in our dataset and log10 of 0 becomes infinity which is difficult to handle. Adding one removes that possibility and gives us a better set.

```
#Log Transformation of Variables
new.sat$PRICE.use=log10(new.sat$PRICE+1)
new.sat$QUANTITY.use=log10(new.sat$QUANTITY+1)
new.sat$RAINFALL.use = log10(new.sat$RAINFALL+1)
new.sat$MainProduct_Tonne.use=log10(new.sat$CropArea_Hectare+1)
new.sat$CropArea_Hectare.use=log10(new.sat$CropArea_Hectare+1)
new.sat$FarmLabour_Hours.use=log10(new.sat$FarmLabour_Hours+1)
new.sat$FarmLabour_Cost.use=log10(new.sat$FarmLabour_Cost+1)
new.sat$AnimalOwned_Hours.use=log10(new.sat$AnimalOwned_Hours+1)
new.sat$AnimalOwned_Cost.use=log10(new.sat$AnimalOwned_Cost+1)
new.sat$Machinery_Hours.use=log10(new.sat$Machinery_Hours+1)
new.sat$Machinery_Cost.use=log10(new.sat$Machinery_Cost+1)
new.sat$SeedQuantity_Kg.use=log10(new.sat$SeedQuantity_Kg+1)
new.sat$Seed_Cost.use=log10(new.sat$Seed_Cost+1)
new.sat$FertilizerQuantity_Kg.use=log10(new.sat$FertilizerQuantity_Kg+1)
new.sat$Fertilizer_Cost.use=log10(new.sat$Fertilizer_Cost+1)
new.sat$Manure_Tonne.use = log10(new.sat$Manure_Tonne+1)
new.sat$Manure_Cost.use= log10(new.sat$Manure_Cost+1)
new.sat$Insectiside_Cost.use = log10(new.sat$Insectiside_Cost+1)
new.sat$TotalCapital_Cost.use = log10(new.sat$TotalCapital_Cost+1)
```

PARTITION THE DATASET INTO TRAINING AND TEST SET

We partitioned the Dataset into the training and test set using the below code which basically divided the dataset into 75% training dataset and 25% test dataset.

```
#Creating Test and Training Set
## 75% of the sample size
smp_size <- floor(0.75 * nrow(new.sat))
## set the seed to make your partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(new.sat)), size = smp_size)

train <- new.sat[train_ind, ]
test <- new.sat[-train_ind, ]
```

HANDLING NA VALUES

```
#Set NA values to 0
train$Cereals[is.na(train$Cereals)] <- 0
```

MULTIPLE LINEAR REGRESSION MODEL

MODEL 1

We build the below model using multiple linear regression:

Price of a crop = β_0 +β_1* **Group of crop** +β_2 * **Quantity of Arrival in market** + β_3 * **Rainfall** + β_4 * **Main Product** + β_5 * **Farm Labor Hours** + β_6 * **Farm Labor Cost** + β_7 * **Animal Owned Hours** + β_8 * **Animal Owned Cost** + β_9 * **Machinery Hours** + β_{10} * **Machinery Cost** + β_{11} * **Fertilizer Quantity** + β_{12} * **Fertilizer Cost** + β_{13} * **Manure Quantity** + β_{14} * **Manure Cost** + β_{15} * **Insecticide Cost** + β_{16} * **Total Capital Cost** +Є

```
#Model1
fit.train <- lm(train$PRICE.use ~
                  train$Cereals+
                  train$Drug.and.Narcotics+
                  train$Dry.Fruits+
                  train$Fibre.Crops+
                  train$Flowers+
                  train$Forest.Products+
                  train$Fruits+
                  train$Live.Stock.Poultry.Fisheries+
                  train$Oil.Seeds+
                  train$Other+
                  train$Pulses+
                  train$Spices+
                  train$Vegetables+
                  train$QUANTITY.use+
                  train$RAINFALL.use+
                  train$FarmLabour_Hours.use+
                  train$FarmLabour_Cost.use+
                  train$Insectiside_Cost.use+
                  train$MainProduct_Tonne.use+
                  train$TotalCapital_Cost.use+
                  train$Manure_Cost.use+
                  train$Manure_Tonne.use+
                  train$Fertilizer_Cost.use+
                  train$FertilizerQuantity_Kg.use+
                  train$Seed_Cost.use+
                  train$SeedQuantity_Kg.use+
                  train$Machinery_Cost.use+
                  train$Machinery_Hours.use+
                  train$AnimalOwned_Cost.use+
                  train$AnimalOwned_Hours.use+
                  train$CropArea_Hectare.use
                , data=train)

summary(fit.train)


Coefficients: (2 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         1.026148   0.046629  22.007  < 2e-16 ***
train$Cereals                       0.032727   0.002514  13.020  < 2e-16 ***
train$Drug.and.Narcotics            0.811234   0.005344 151.804  < 2e-16 ***
train$Dry.Fruits                    0.812874   0.014370  56.565  < 2e-16 ***
train$Fibre.Crops                   0.471324   0.003699 127.408  < 2e-16 ***
train$Flowers                      -0.225698   0.037861  -5.961 2.51e-09 ***
train$Forest.Products              -0.008503   0.005854  -1.452 0.146392
train$Fruits                        0.183312   0.003302  55.523  < 2e-16 ***
train$Live.Stock.Poultry.Fisheries  0.789181   0.207053   3.811 0.000138 ***
train$Oil.Seeds                     0.399807   0.002779 143.863  < 2e-16 ***
train$Other                         0.410699   0.003847 106.766  < 2e-16 ***
train$Pulses                        0.414306   0.002823 146.741  < 2e-16 ***
train$Spices                        0.645883   0.003274 197.284  < 2e-16 ***
```

```
train$QUANTITY.use              0.022530   0.001004  22.439  < 2e-16 ***
train$RAINFALL.use              0.006500   0.000849   7.656 1.94e-14 ***
train$FarmLabour_Hours.use     -0.259642   0.016263 -15.966  < 2e-16 ***
train$FarmLabour_Cost.use       0.358683   0.013189  27.197  < 2e-16 ***
train$Insectiside_Cost.use      0.029124   0.002905  10.027  < 2e-16 ***
train$MainProduct_Tonne.use     0.044050   0.023722   1.857 0.063326 .
train$TotalCapital_Cost.use     0.006027   0.007681   0.785 0.432692
train$Manure_Cost.use           0.001305   0.004643   0.281 0.778578
train$Manure_Tonne.use          0.004123   0.007848   0.525 0.599321
train$Fertilizer_Cost.use       0.174325   0.024378   7.151 8.70e-13 ***
train$FertilizerQuantity_Kg.use -0.283519  0.030488  -9.299  < 2e-16 ***
train$Seed_Cost.use             0.090301   0.005572  16.205  < 2e-16 ***
train$SeedQuantity_Kg.use      -0.062313   0.004135 -15.071  < 2e-16 ***
train$Machinery_Cost.use       -0.036309   0.003802  -9.550  < 2e-16 ***
train$Machinery_Hours.use      -0.035706   0.009509  -3.755 0.000173 ***
train$AnimalOwned_Cost.use      0.084149   0.008451   9.958  < 2e-16 ***
train$AnimalOwned_Hours.use    -0.146888   0.011739 -12.513  < 2e-16 ***
train$CropArea_Hectare.use            NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.207 on 77047 degrees of freedom
Multiple R-squared:  0.5706,    Adjusted R-squared:  0.5705
F-statistic:  3531 on 29 and 77047 DF,  p-value: < 2.2e-16
```

MODEL 1: PREDICTION ACCURACY AND SUMMARY

```
predTst <- predict(fit.train, test, interval="prediction")
summary(predTst)

      fit             lwr             upr
 Min.   :1.544   Min.   :1.131   Min.   :1.956
 1st Qu.:1.909   1st Qu.:1.504   1st Qu.:2.315
 Median :2.197   Median :1.791   Median :2.603
 Mean   :2.140   Mean   :1.734   Mean   :2.546
 3rd Qu.:2.308   3rd Qu.:1.902   3rd Qu.:2.713
 Max.   :2.864   Max.   :2.458   Max.   :3.274
```

```
rmse <- function(error)
{
  sqrt(mean(error^2))
}

rmse(fit.train$residuals)
```

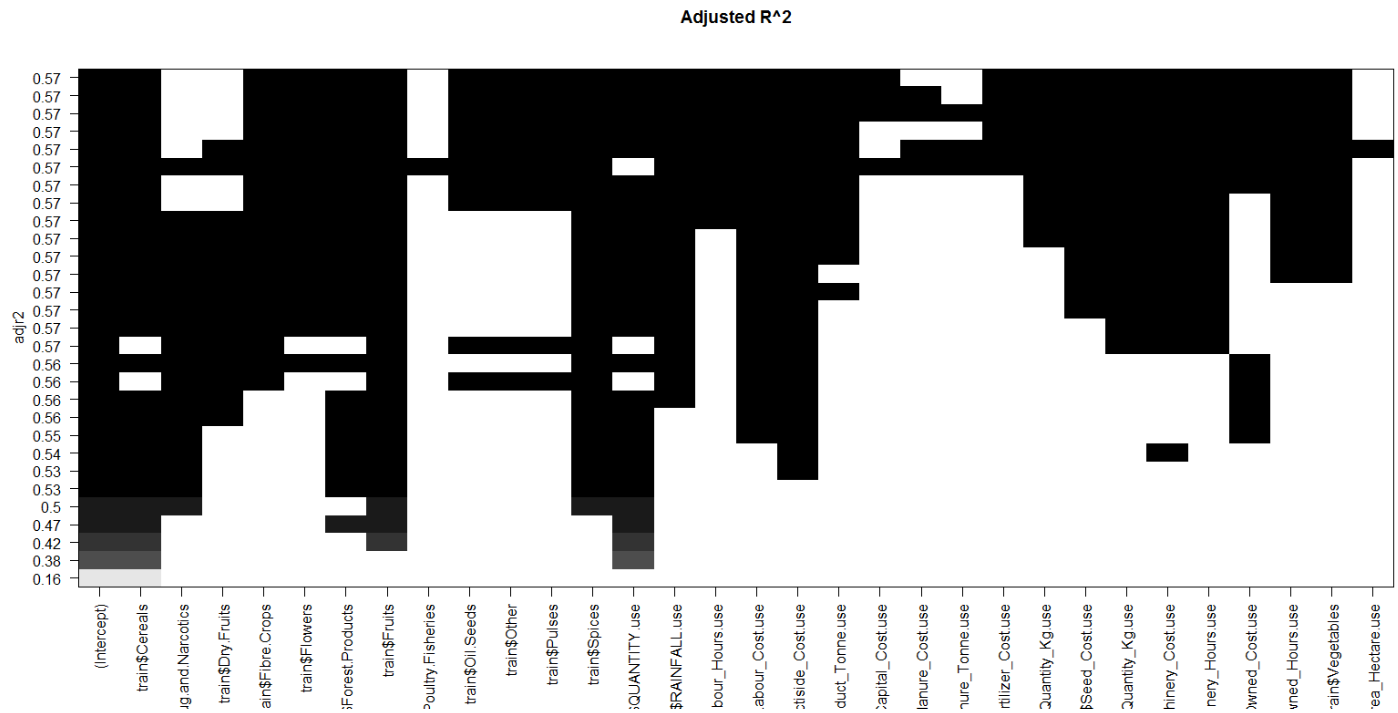
```
> rmse(fit.train$residuals)
[1] 0.206989
```

Above is the result of the linear regression model on the test data which predicts the lower bound, upper bound and the fit value for the model. We also have predicted the RMSE value of the model which comes out to be 0.206989 which is high in price predictions.

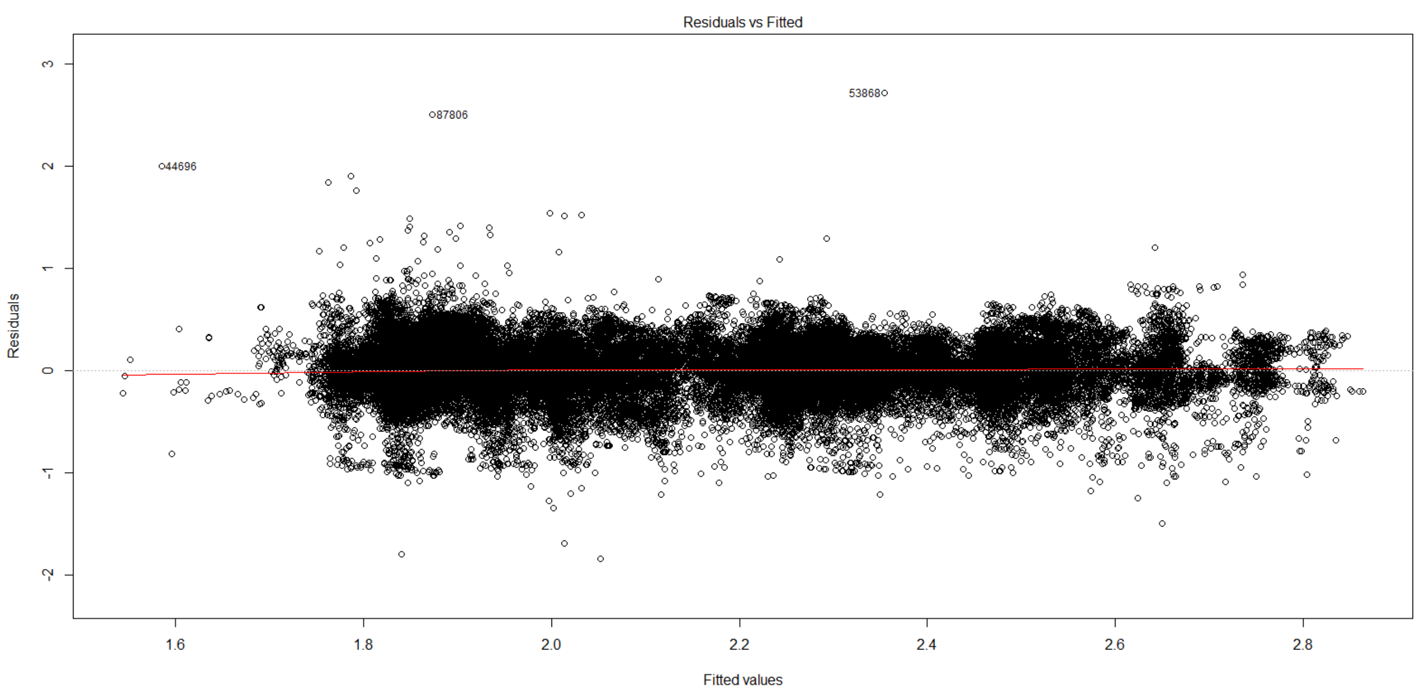
MODEL 1: INTERPRETATION OF THE REGRESSION OUTPUT

- The intercept value(β_0) is 1.0268.
- Most of the variables are all statistically significant.
- The multiple R^2 which is the percentage of variation in the response variable that is explained by the variations in the explanatory variables is rather good at about 0.5706.
- The adjusted R^2 is almost same as R^2
- The RMSE for the model is 0.2069 which means that the multiple linear regression model has an accuracy percentage of 20.69 %.

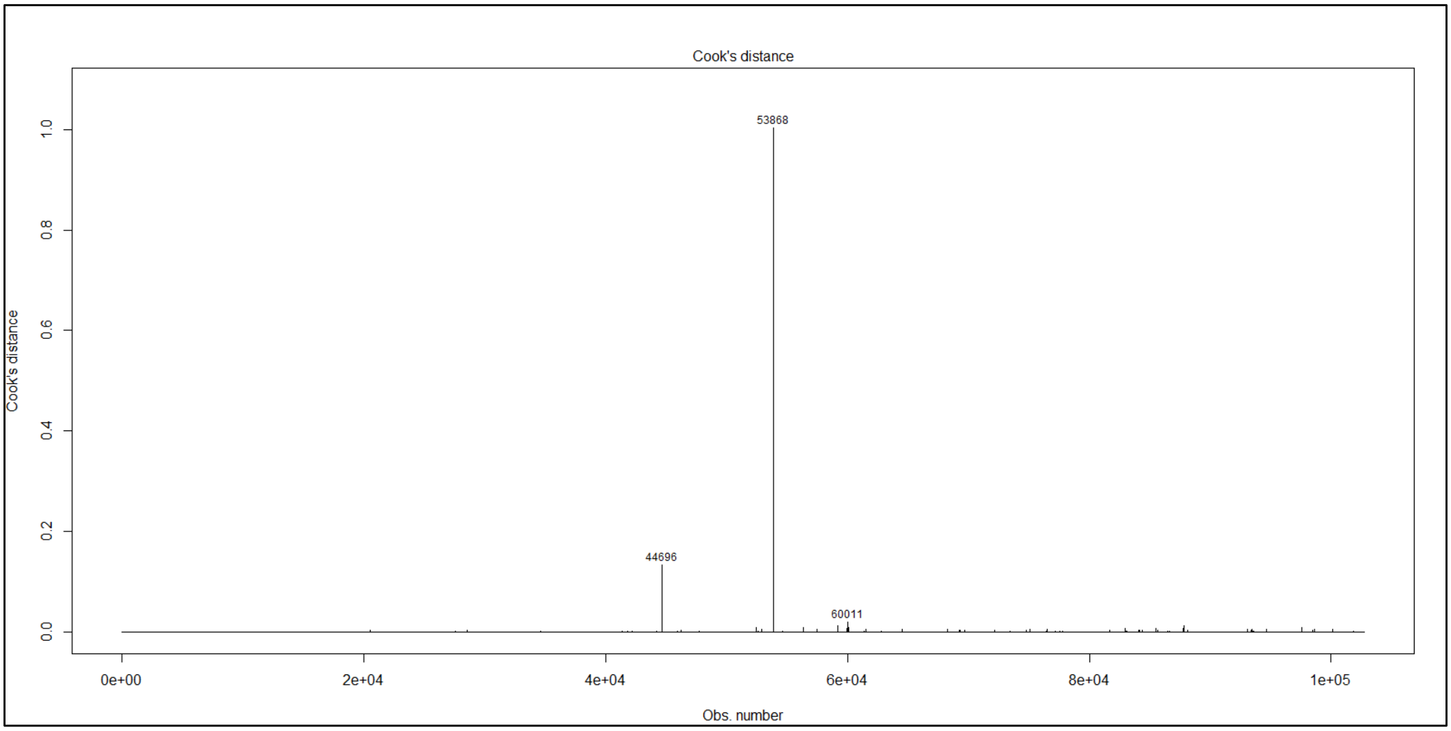
MODEL 1: MULTIPLE LINEAR REGRESSION REGRESSION SUBSET PLOT



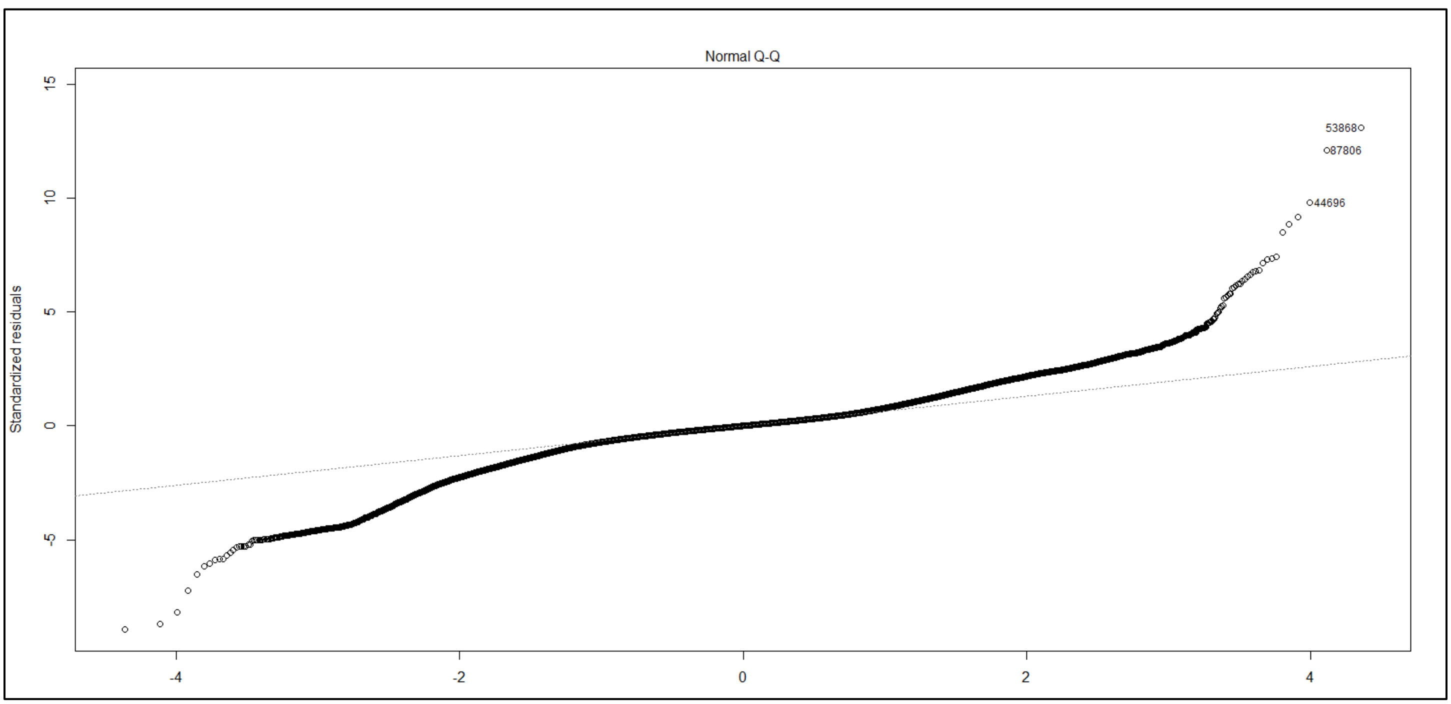
MODEL 1: RESIDUALS VS FITTED VALUE PLOT



MODEL 1: FINDING THE INFLUENTIAL OBSERVATIONS USING COOKS DISTANCE



MODEL 1: Q-Q PLOT OF RESIDUALS



PRICE PREDICTION USING TIME SERIES FORECASTING

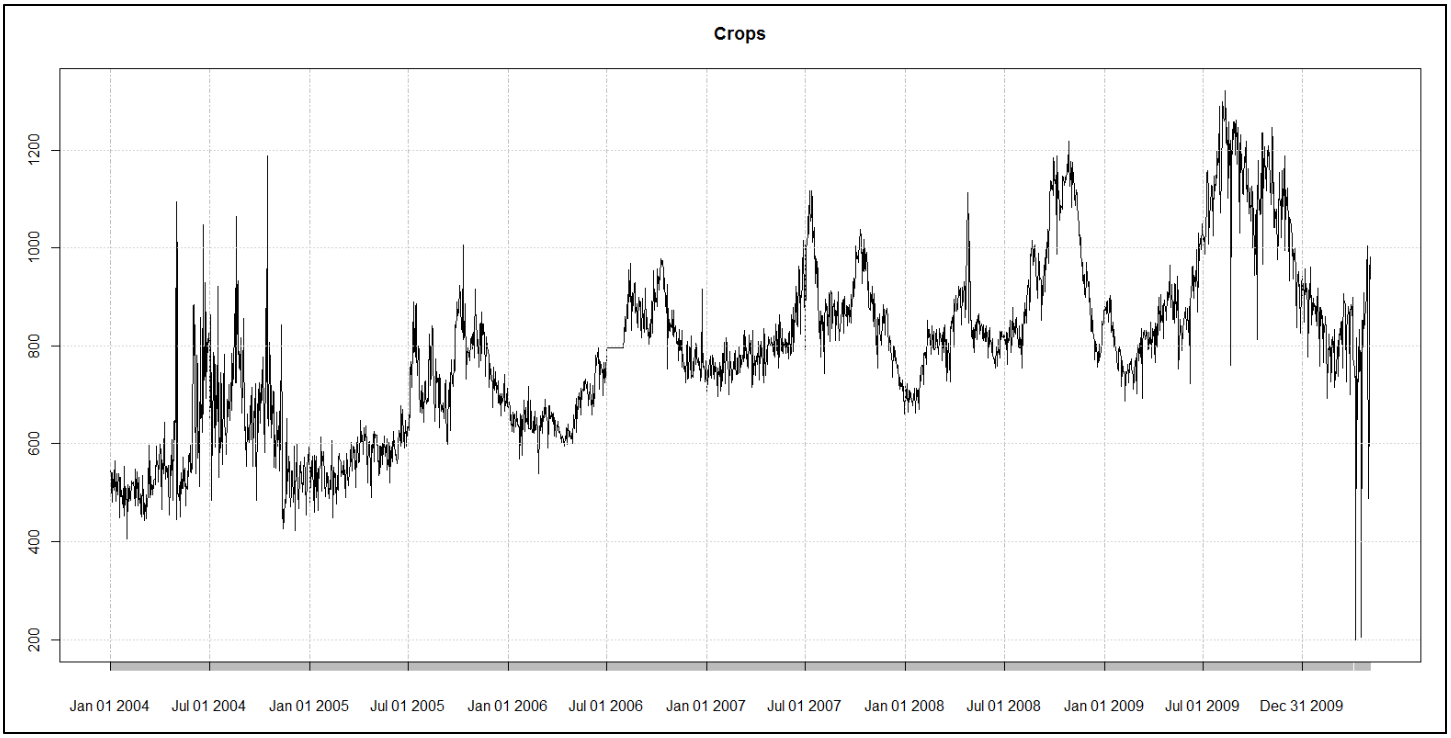
BASIC STATISTICS OF DATA TO BE USED FOR TIME SERIES FORECASTING

In this model we try to work with the Mandi Prices Dataset and try to construct a time series forecasting model at a grain of a day level.

Below we point the basic statistics of the data at hand.

```
> start(CropPrices)
[1] "2004-01-01"

> end(CropPrices)
[1] "2010-05-05"
> frequency(CropPrices)
[1] 1
> summary(CropPrices)
     Index              CropPrices
 Min.   :2004-01-01   Min.   :-0.7393334
 1st Qu.:2005-08-01   1st Qu.:-0.0317695
 Median :2007-03-02   Median :-0.0002122
 Mean   :2007-03-02   Mean   : 0.0058404
 3rd Qu.:2008-09-30   3rd Qu.: 0.0333135
 Max.   :2010-05-05   Max.   : 3.0870161
                      NA's   :1
```

VISUALIZATION OF TIME SERIES PLOTS



There seems to be a growing pattern in the time series apart from the lag at latter half. We will try to perform the Time Series forecasting to forecast the future values.

MODEL 1: MEAN METHOD

We wanted to first try a very simple method of forecasting the prices of crop groups in a market in a state using the average method. The forecasts of all the future values are equal to the mean of the historical data.

```
meanf(Data$Price, 20)
```

And here is the forecast we got based on this very simple and elegant model.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1	787.2035	565.7332	1008.674	448.4133	1125.994
2	787.2035	565.7332	1008.674	448.4133	1125.994
3	787.2035	565.7332	1008.674	448.4133	1125.994
4	787.2035	565.7332	1008.674	448.4133	1125.994
5	787.2035	565.7332	1008.674	448.4133	1125.994
6	787.2035	565.7332	1008.674	448.4133	1125.994
7	787.2035	565.7332	1008.674	448.4133	1125.994
8	787.2035	565.7332	1008.674	448.4133	1125.994
9	787.2035	565.7332	1008.674	448.4133	1125.994
10	787.2035	565.7332	1008.674	448.4133	1125.994
11	787.2035	565.7332	1008.674	448.4133	1125.994
12	787.2035	565.7332	1008.674	448.4133	1125.994
13	787.2035	565.7332	1008.674	448.4133	1125.994
14	787.2035	565.7332	1008.674	448.4133	1125.994
15	787.2035	565.7332	1008.674	448.4133	1125.994
16	787.2035	565.7332	1008.674	448.4133	1125.994
17	787.2035	565.7332	1008.674	448.4133	1125.994
18	787.2035	565.7332	1008.674	448.4133	1125.994
19	787.2035	565.7332	1008.674	448.4133	1125.994
20	787.2035	565.7332	1008.674	448.4133	1125.994

MODEL 2: NAÏVE FORECAST METHOD

We also developed another simple yet elegant method, the naïve forecast method. The method says that all observations are simply set to be the value of the last observation.

```
rwf(Data$Price, 20)
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2315	980.9267	892.2499	1069.603	845.3073	1116.546
2316	980.9267	855.5188	1106.334	789.1319	1172.721
2317	980.9267	827.3341	1134.519	746.0271	1215.826
2318	980.9267	803.5732	1158.280	709.6880	1252.165
2319	980.9267	782.6395	1179.214	677.6726	1284.181
2320	980.9267	763.7139	1198.139	648.7285	1313.125
2321	980.9267	746.3101	1215.543	622.1116	1339.742
2322	980.9267	730.1110	1231.742	597.3372	1364.516
2323	980.9267	714.8965	1246.957	574.0686	1387.785
2324	980.9267	700.5062	1261.347	552.0606	1409.793
2325	980.9267	686.8192	1275.034	531.1282	1430.725
2326	980.9267	673.7415	1288.112	511.1274	1450.726
2327	980.9267	661.1982	1300.655	491.9441	1469.909
2328	980.9267	649.1287	1312.725	473.4855	1488.368
2329	980.9267	637.4832	1324.370	455.6752	1506.178
2330	980.9267	626.2198	1335.634	438.4493	1523.404
2331	980.9267	615.3032	1346.550	421.7538	1540.100
2332	980.9267	604.7032	1357.150	405.5425	1556.311
2333	980.9267	594.3938	1367.460	389.7756	1572.078
2334	980.9267	584.3523	1377.501	374.4185	1587.435

We also tried the seasonal naïve method as well as crops could be prone to seasonality:

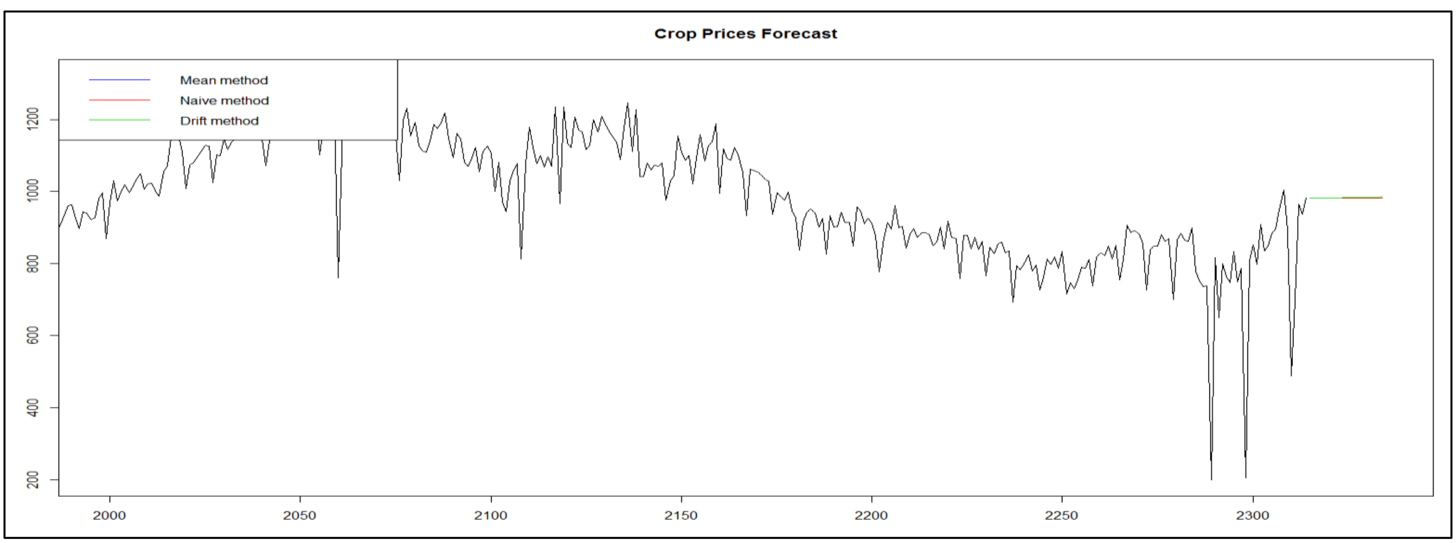
```
snaive(Data$Price, 20)
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2315	980.9267	892.2499	1069.603	845.3073	1116.546
2316	980.9267	855.5188	1106.334	789.1319	1172.721
2317	980.9267	827.3341	1134.519	746.0271	1215.826
2318	980.9267	803.5732	1158.280	709.6880	1252.165
2319	980.9267	782.6395	1179.214	677.6726	1284.181
2320	980.9267	763.7139	1198.139	648.7285	1313.125
2321	980.9267	746.3101	1215.543	622.1116	1339.742
2322	980.9267	730.1110	1231.742	597.3372	1364.516
2323	980.9267	714.8965	1246.957	574.0686	1387.785
2324	980.9267	700.5062	1261.347	552.0606	1409.793
2325	980.9267	686.8192	1275.034	531.1282	1430.725
2326	980.9267	673.7415	1288.112	511.1274	1450.726
2327	980.9267	661.1982	1300.655	491.9441	1469.909
2328	980.9267	649.1287	1312.725	473.4855	1488.368
2329	980.9267	637.4832	1324.370	455.6752	1506.178
2330	980.9267	626.2198	1335.634	438.4493	1523.404
2331	980.9267	615.3032	1346.550	421.7538	1540.100
2332	980.9267	604.7032	1357.150	405.5425	1556.311
2333	980.9267	594.3938	1367.460	389.7756	1572.078
2334	980.9267	584.3523	1377.501	374.4185	1587.435

We also tried the DRIFT method which allows the forecast to increase or decrease where the amount of change over time (called the drift) is set to be the average change seen in the historical data.

```
fit=rwf(Data$Price, 20, drift=TRUE)
```

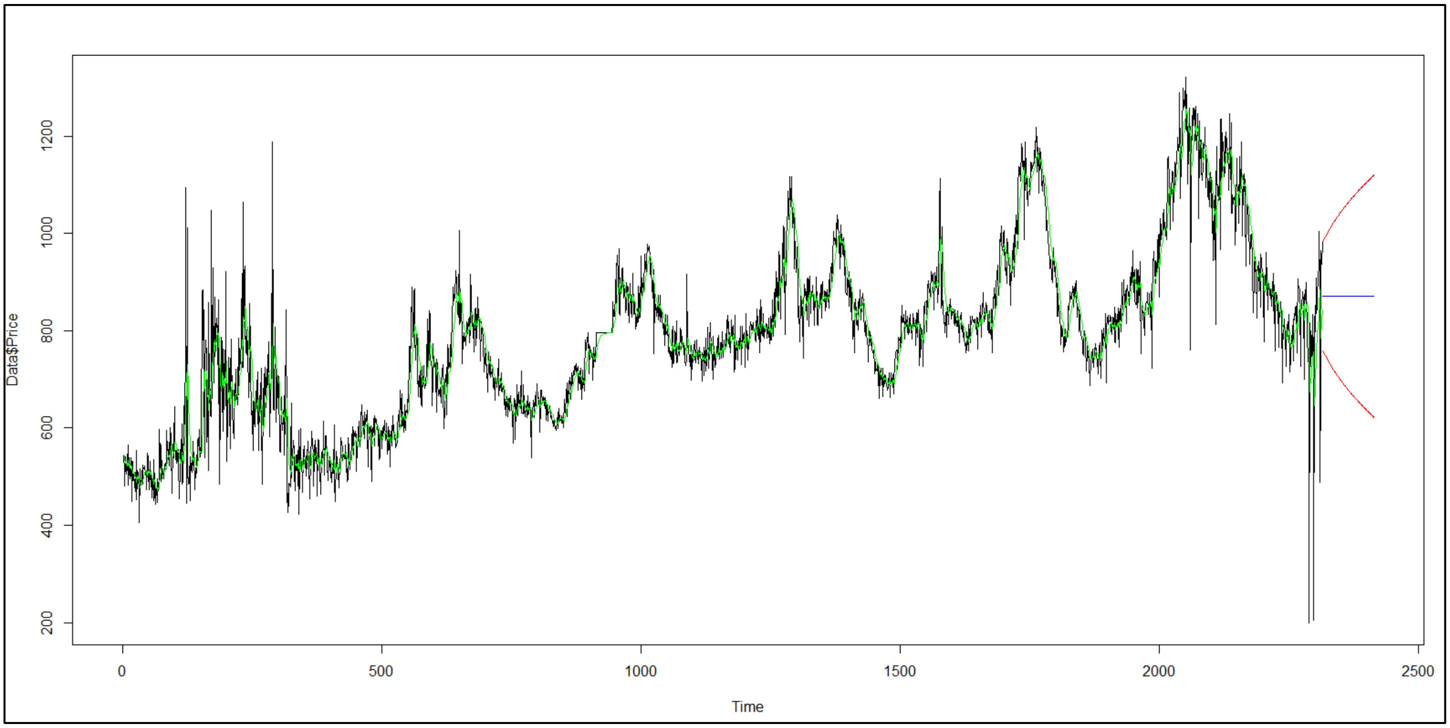
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2315	981.1152	892.4196	1069.811	845.4670	1116.763
2316	981.3037	855.8421	1106.765	789.4267	1173.181
2317	981.4922	827.8005	1135.184	746.4411	1216.543
2318	981.6807	804.1745	1159.187	710.2084	1253.153
2319	981.8692	783.3684	1180.370	678.2885	1285.450
2320	982.0577	764.5641	1199.551	649.4299	1314.685
2321	982.2462	747.2756	1217.217	622.8897	1341.603
2322	982.4347	731.1864	1233.683	598.1836	1366.686
2323	982.6232	716.0767	1249.170	574.9755	1390.271
2324	982.8117	701.7865	1263.837	553.0208	1412.603
2325	983.0002	688.1950	1277.805	532.1346	1433.866
2326	983.1887	675.2085	1291.169	512.1736	1454.204
2327	983.3772	662.7523	1304.002	493.0237	1473.731
2328	983.5657	650.7659	1316.365	474.5924	1492.539
2329	983.7542	639.1997	1328.309	456.8035	1510.705
2330	983.9427	628.0119	1339.874	439.5935	1528.292
2331	984.1312	617.1673	1351.095	422.9084	1545.354
2332	984.3197	606.6359	1362.004	406.7022	1561.937
2333	984.5082	596.3918	1372.625	390.9353	1578.081
2334	984.6967	586.4122	1382.981	375.5731	1593.820



MODEL 3: EXPONENTIAL SMOOTHING

We will try to build the model using the holt winters method below

```
# simple exponential - models level
fit.mean <- HoltWinters(x=Data$Price, alpha = 0.2,beta=FALSE, gamma=FALSE)
#Predicting 100 days ahead in time
fit.predict <-predict(fit.mean,n.ahead=100,prediction.interval=TRUE)
fit.predict
#Plot of fitted value upper bound and lower bound
plot.ts(Data$Price,xlim=c(0,2414))
lines(fit.mean$fitted[,1],col="green")
lines(fit.predict[,1],col="blue")
lines(fit.predict[,2],col="red")
lines(fit.predict[,3],col="red")
```



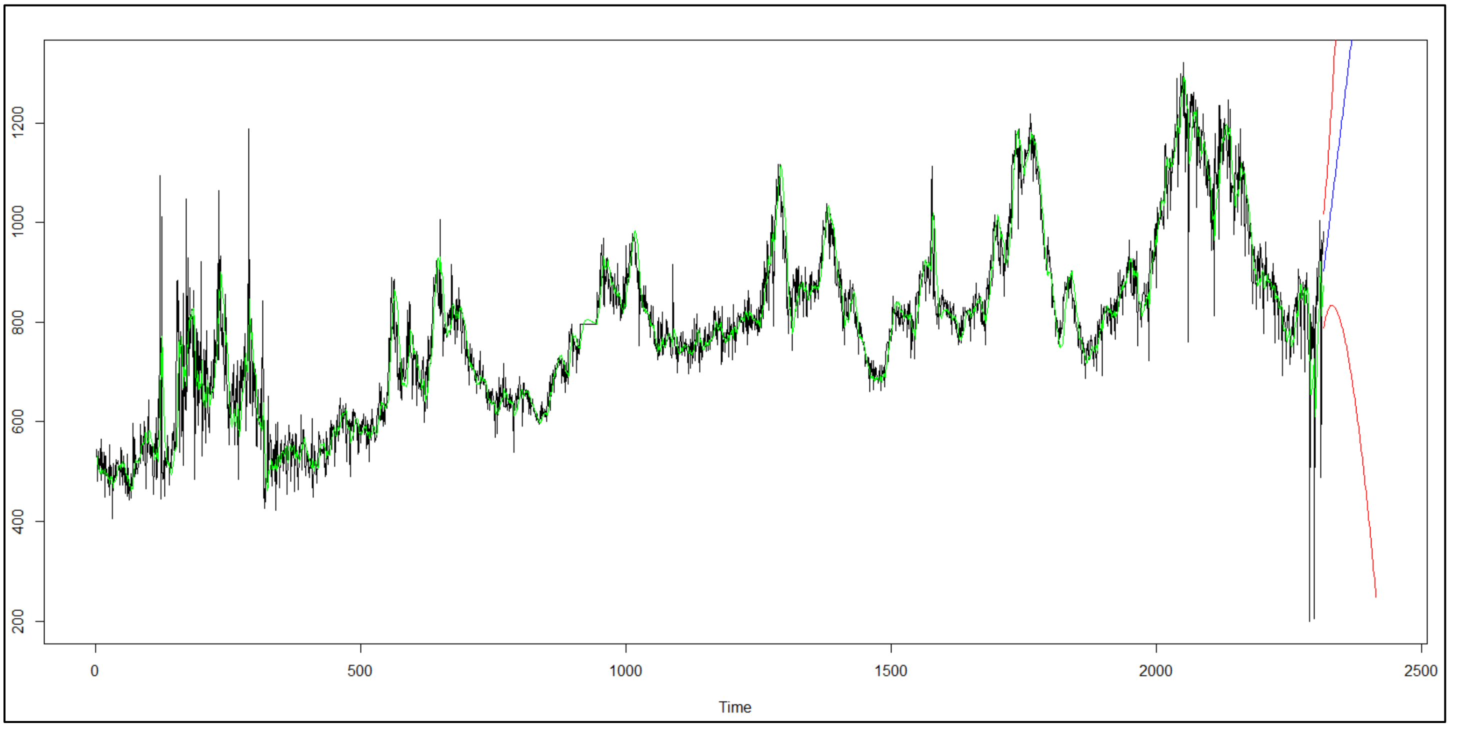
The green line shows exponentially weighted moving average of the fitted value as we chose alpha = 0.2, the blue line shows the prediction and the red lines show the upper and lower bound of the prediction.

Here below are the predicted values.

	fit	upr	lwr
2315	871.03	982.7992	759.2608
2316	871.03	985.0127	757.0473
2317	871.03	987.1840	754.8760
2318	871.03	989.3154	752.7446
2319	871.03	991.4091	750.6509
2320	871.03	993.4671	748.5930
2321	871.03	995.4909	746.5691
2322	871.03	997.4825	744.5776
2323	871.03	999.4431	742.6170
2324	871.03	1001.3742	740.6858
2325	871.03	1003.2771	738.7829
2326	871.03	1005.1531	736.9070
2327	871.03	1007.0031	735.0569
2328	871.03	1008.8284	733.2317

We now try to add a trend to the model by setting the beta.

```
# simple exponential - models level
fit.mean <- HoltWinters(x=Data$Price, alpha = 0.2,beta=0.1, gamma=FALSE)
#Predicting 100 days ahead in time
fit.predict <-predict(fit.mean,n.ahead=100,prediction.interval=TRUE)
fit.predict
#Plot of fitted value upper bound and lower bound
plot.ts(Data$Price,xlim=c(0,2414))
lines(fit.mean$fitted[,1],col="green")
lines(fit.predict[,1],col="blue")
lines(fit.predict[,2],col="red")
lines(fit.predict[,3],col="red")
```



Now we see that it has added a trend which is going upwards. This will help in predicting the trend of the prices. The predicted data and trend looks like below.

	fit	upr	lwr
2315	902.8731	1017.639	788.1068
2316	911.6374	1029.148	794.1266
2317	920.4017	1041.097	799.7060
2318	929.1660	1053.496	804.8365
2319	937.9303	1066.346	809.5152
2320	946.6947	1079.645	813.7440
2321	955.4590	1093.389	817.5292
2322	964.2233	1107.566	820.8803
2323	972.9876	1122.166	823.8091
2324	981.7519	1137.174	826.3294
2325	990.5162	1152.577	828.4559
2326	999.2805	1168.357	830.2037
2327	1008.0449	1184.501	831.5883
2328	1016.8092	1200.994	832.6245
2329	1025.5735	1217.820	833.3272
2330	1034.3378	1234.965	833.7101
2331	1043.1021	1252.418	833.7866
2332	1051.8664	1270.164	833.5691
2333	1060.6307	1288.192	833.0692
2334	1069.3951	1306.492	832.2979

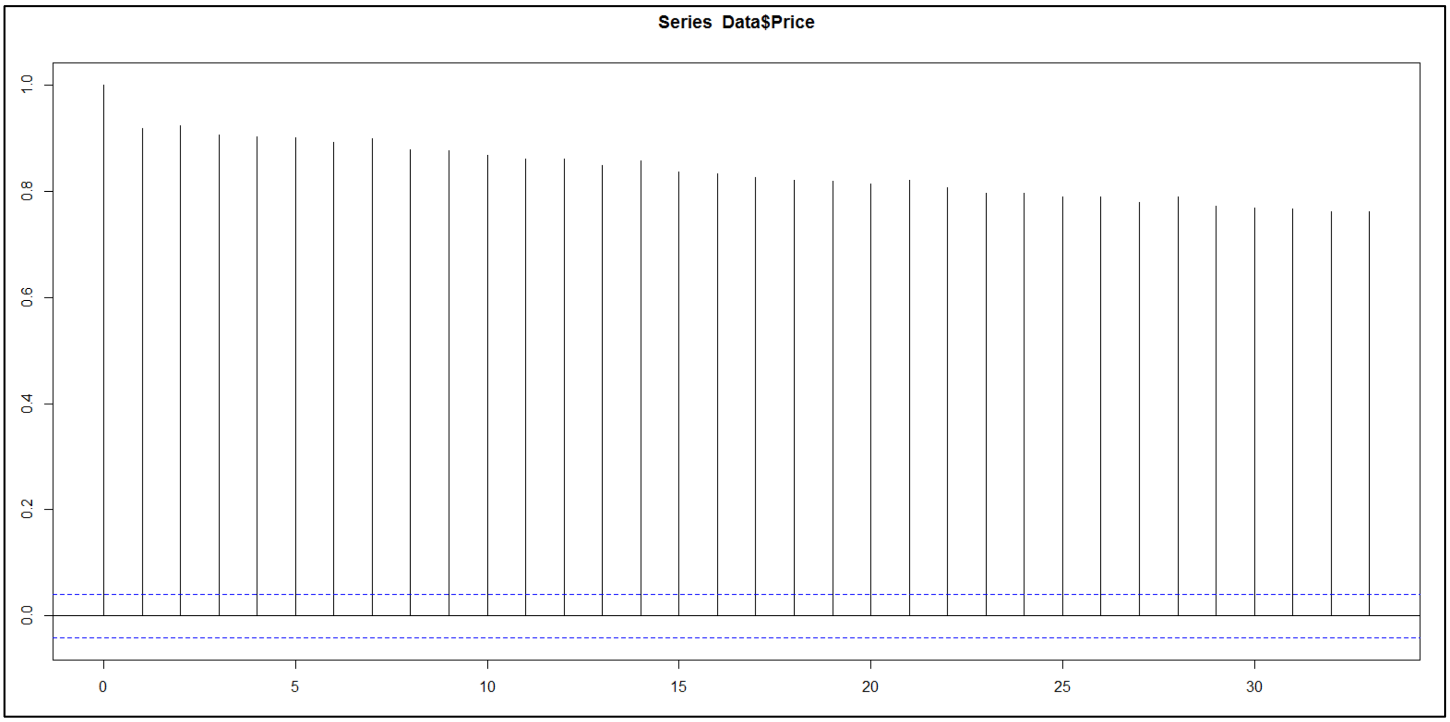
MODEL 4: ARIMA MODEL

We also build an ARIMA based model to predict the price of a crop group in a market in a State. We are going to determine the number of AR and MA terms using AUTO.ARIMA function.

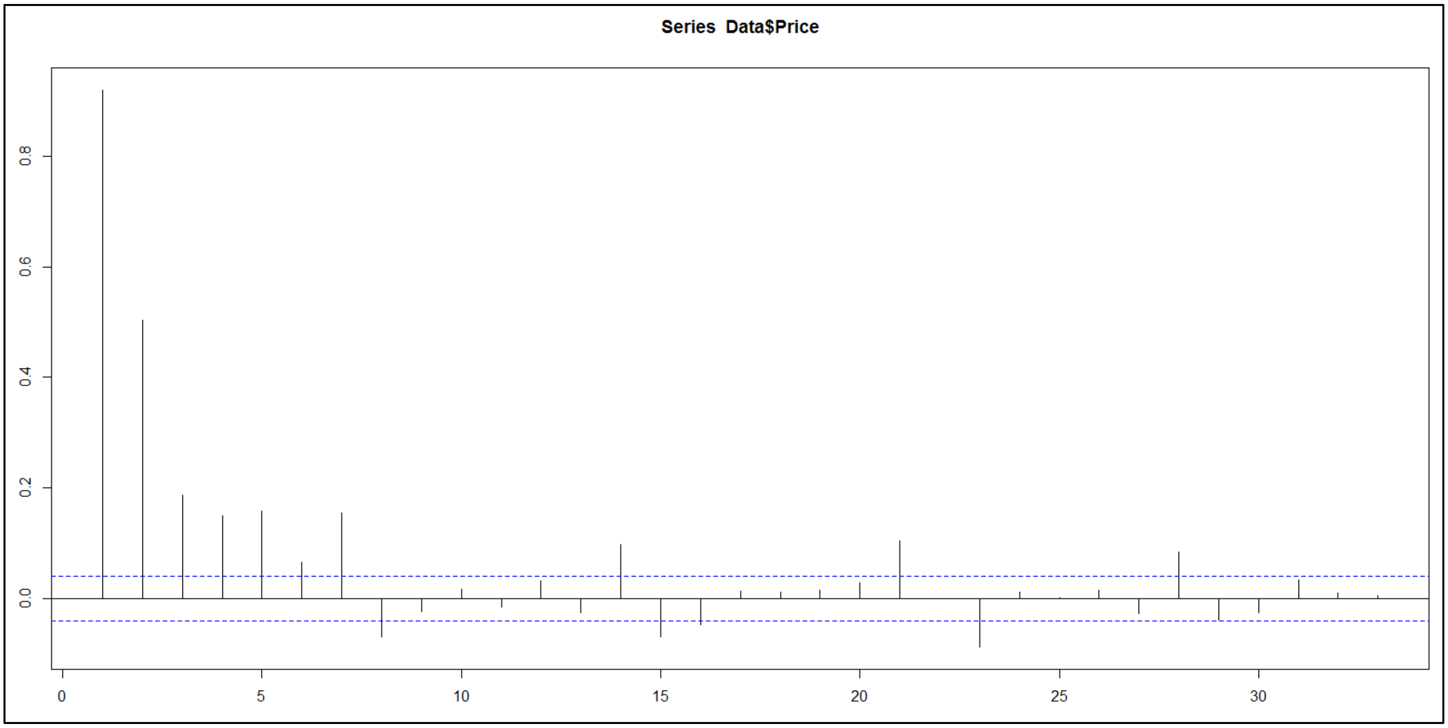
```
Coefficients:
          ar1      ar2      ar3      ar4     ma1      ma2
      -0.7828  -0.0097  -0.0522  -0.1194  0.0687  -0.4410
s.e.   0.0768   0.0685   0.0496   0.0264  0.0755   0.0602

sigma^2 estimated as 3090:  log likelihood=-12572.81
AIC=25159.62   AICc=25159.66   BIC=25199.84
```

The output of AUTO.ARIMA tells us that there are total of 4 AR terms and 2 MA terms.

ACF PLOT



PACF PLOT



We then tried to build the ARIMA model using these features.

```
acf(Data$Price)
pacf(Data$Price)
auto.arima(Data$Price)
fit1<- arima(Data$Price, order = c(4,0,2))
```

DIAGNOSTIC TESTS USING LJUNG BOX TEST

We performed the diagnostic tests on the model using the LJung-Box test.

```
fit1_resid <-residuals(fit1)
Box.test(fit1_resid, lag=10, type = "Ljung-Box")


        Box-Ljung test

data:  fit1_resid
X-squared = 58.832, df = 10, p-value = 6.024e-09
```

But since the p value is very low for this model the model is not so good at prediction of the price of the crop.

MANDI PRICE PREDICTOR MOBILE APPLICATION

We have formulated an idea to develop a mobile application which will help both farmers and traders get a price prediction of various crop groups in their local markets. The app will be powered by an API which will continuously be feed by the Mandi price market data, rainfall data and the cost of cultivation data. We present below the wireframes of the mobile application:

WIREFRAMES OF MOBILE APPLICATION

The app can be easily downloaded from play store. The is using a symbol of swastika on a crop field which is very auspicious. The UI is kept extremely simple so that any individual with the very basic knowledge of how to operate a mobile phone can use this application. The application ask user to provide the State and the Market name and will in turn provide the Prices for all the Crop groups in that market. It will also provide an indicator which will suggest if the price has moved up or down month on month in that market.