Pricing pattern determination of Organic versus Conventional Produce (Fruits and vegetables).



A Capstone project:

Vidya Ramasamy

Under the mentorship:

Dr. Marko Mitic (Data Scientist at Telenor, Belgrade, Serbia)

For the course:

Foundations of Data-science (Springboard)

Table of Contents:

1. Introduction………………………………………………………………….............3
2. Purpose…………………………………………………………………………………….3
3. Data source……………………………………………………………………………….4
4. Data preparation……………………………………………………………………....5
5. Modeling and Data Analysis………………………………………………………6
6. Descriptive Analysis……………………………………………………………….….8
7. Exploratory Analysis………………………………………………………………….9
8. Predictive Analysis
   1. Linear regression and Anova modeling ………………………….10
   2. K mean clustering ………………………………………………………… 20
   3. Location analysis ……………………………………………………………21
   4. Organic and conventional price differentiation………………23
   5. Month-wise pricing analysis……………………………………………25

9.Evaluation and Deployment …………………………………………………………….31

10. References……………………………………………………………………………………….33



# INTRODUCTION

Food has been an important attribute of our daily lives. The type of our food intake will have a direct impact on our health. Hence, my study is based on market price pattern of Organic versus Conventional produces (fruits and Vegetables only) across two different geographical locations in USA.

This study considers data from **San Francisco** (west coast) and **Atlanta** (east coast). The type of market explored in this study is Terminal market or Wholesale market (a logistics Hub for distribution of produce).

**In this study, I want to solve following problems:**

1. What are the **price ranges** of organic vegetables vs Conventional vegetables?
2. What is the **probability of the increase** in price using predictive Analysis?
3. Which **market location has more price difference**?
4. Which **vegetable/fruit is expensive**?

# Purpose

The purpose of this project:

* 1. Understand the organic market prices in two terminal markets.
  2. The price differentiation between conventional prices and organic prices (organic versus conventional)
  3. The consumers will know how to purchase better with data analysis findings. If consumers learn about the price comparatively, then they will make better decisions based on the analysis. This project gives a monthly breakdown in comparison on which will help in being selective on procurement of produce
  4. Is there a correlation in prices between two different geographical locations? (SFO and ATL)

My client will be **Individual Wholesale Retailers/Consumers** who are interested in buying Organic and conventional produce.

The List of Consumers are:

1. Restaurant/Hotel
2. Grocery Retailers
3. Individuals /Home
4. Culinary schools

Below are the list of wholesale Consumers:

<http://www.sfproduce.org/merchants/merchname.html>

<http://terminalmarkets.com/atlantastate.htm>

Outcome of Statistical Analysis:

1. Pattern in produce pricing – A comparison of Organic and conventional Produce
2. Do the Geo-graphical locations differ in their pricing pattern? (San Francisco and Atlanta)
3. (consumers can be selective on the seasonal availability/Month=wise pricing) example: Restaurant owners can decide which produce to buy according to the Price increase-decrease and their operating cost.
4. Predictive analysis of prices will help the consumer on the produce prices that are probable to increase
5. Get perspectives on Organic produce, purchasing organic versus conventional in the long run.

# Data-Source

I extracted the dataset from “United States Department of Agriculture”

Website: <http://www.ers.usda.gov/data-products/organic-prices.aspx>



# **4.**Data Preparation

Data wrangling of the dataset:

Data Analysis- Understood the type of dataset

1. Categorical values: Commodity/Category/Location
2. Numerical values: January\_2012 prices to December\_2013
3. Data was stored as Excel files

Extraction - excel converted as csv file

TRANFORM – Replace Missing values with median value prices

LOAD – TMKTPRICE DATASET IN R STUDIO

R CODE:

* File path : setwd("C:/Users/vidya/Desktop/datasets")
* Storage of dataset in R studio: The csv file is stored as “veg” dataframe.

veg <- read.csv("tmktprice.csv", stringsAsFactors = FALSE)

* To understand the structure of the dataset(categorical/character/numerical)

str(veg)

* To view the table in R studio

View(veg)

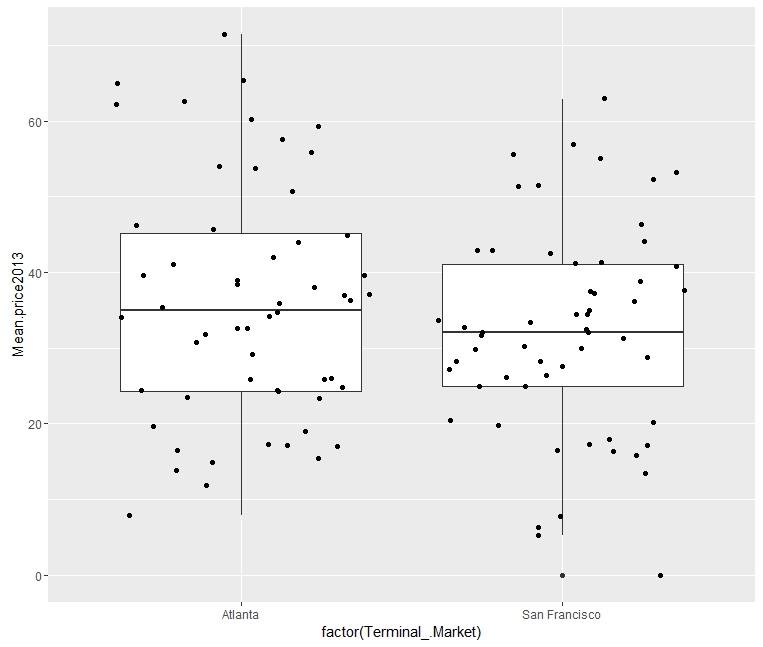
# **5.**Modeling and Data Analysis

Statistical Methodology:

1. Descriptive Analysis
2. Exploratory Analysis
3. Predictive Analysis

A basic plot that depicts the Average price pattern of produce in 2013 in two different locations (San Francisco and Atlanta) This type of analysis was done to understand the dataset with

the average median price. It depicts the basic price pattern of the produce.



This simple Box plot reflect the average price pattern; the average median price is higher at Atlanta than San Francisco.

The scattered points depict the prices of various commodities (vegetables and fruits)

Packaging is an important aspect of selling Produce in terminal markets.

Predominantly they are packages in two ways for the different commodities.

A simple Bar graph shows that the type of packaging, predominately are the cartons.



Flat- A container that is packed in 1 or 2 layers; often used to pack produce that is already packaged in consumer-ready containers, like berries

Carton -A corrugated fiberboard container that closes with top flaps. It ranges in size from ½ peck to 1 bushel and is often waxed for water resistance and to prevent containers from falling apart if they become wet or moist

Flat Carton

# 6.Descriptive Analysis

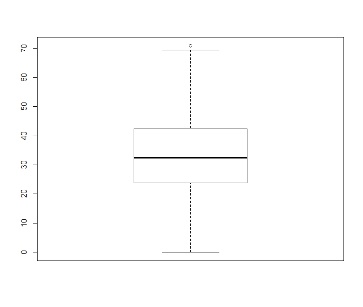
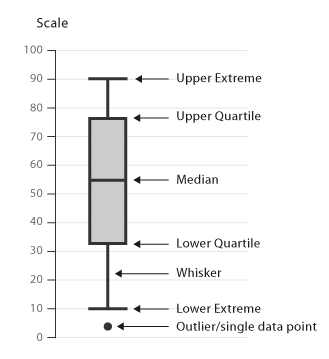
Descriptive Analysis is the basic analysis of a given column:

|  |  |
| --- | --- |
| Month and Year | (Raw)Average Price of Produce at Terminal Market |
| January 2012 | 31.15368 |
| January 2013 | 34.84596 |
| February 2012 | 30.88149 |
| February 2013 | 33.98053 |
| March 2012 | 31.49246 |
| March 2013 | 33.48377 |
| April 2012 | 31.83675 |
| April 2013 | 32.9964 |
| May 2012 | 31.39912 |
| May 2013 | 32.56193 |
| June 2012 | 31.88289 |
| June 2013 | 33.38044 |
| July 2012 | 32.48763 |
| July 2013 | 33.91561 |
| August 2012 | 33.81404 |
| September 2012 | 34.75509 |
| September 2013 | 34.16254 |
| October 2012 | 33.46395 |
| October 2013 | 35.21175 |
| November 2012 | 32.47614 |
| November 2013 | 34.39789 |
| December 2012 | 32.75009 |
| December 2013 | 33.70904 |

Example: In December 2013, The average price of a produce was $33.71 and maximum price was $71.00. The Image depicts the price pattern at “Upper extreme” and “lower extreme”

Column name: December\_2013

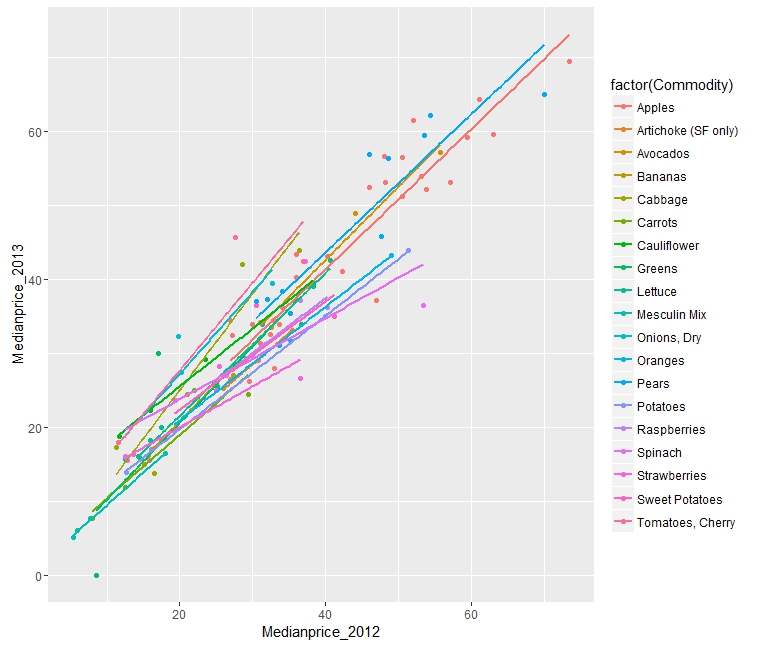
Min: 0.00 Low.Qu:24.01 Median:32.49 Mean:33.71 Upper.Qu:42.12 Max:71.00

# **7**.Exploratory Analysis

The dataset needs to be explored. The structure of the dataset, levels, columns and rows that exist in the dataset.

There are 19 types of produce (Vegetables and fruits) categorized into organic and conventional. Their corresponding prices for consecutive 2 years with their unique corresponding average prices.

Scatter-plot of different commodities: 

# Predictive Analysis

* 1. Linear regression and Anova modeling
  2. K mean clustering
  3. Location analysis
  4. Organic and conventional price differentiation
  5. Month-wise pricing analysis

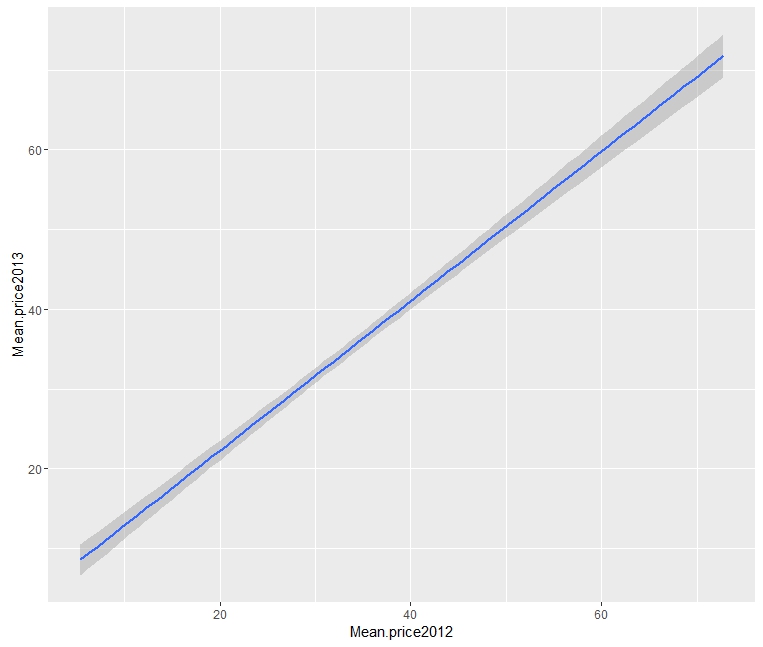
**I. Linear regression and Anova Models:**

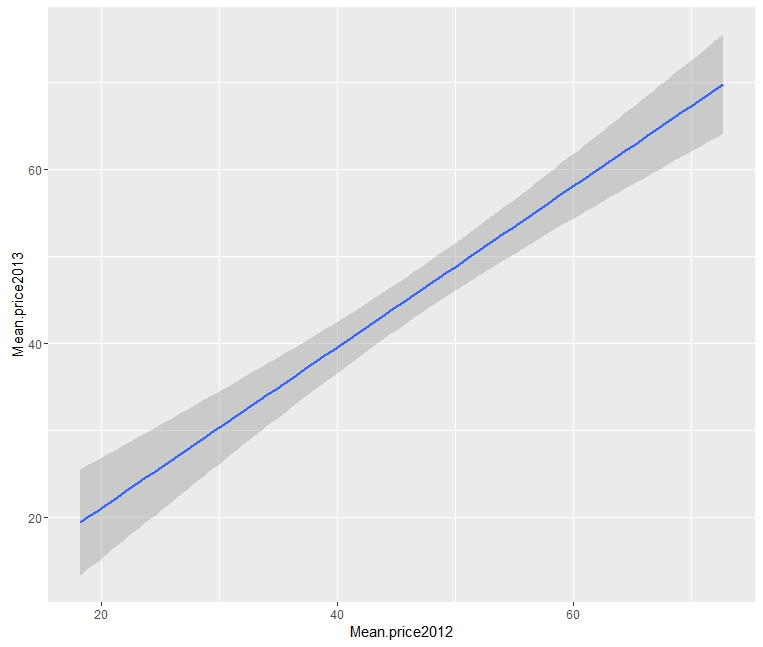
Statistical Analysis:

* 1. Linear Regression Model

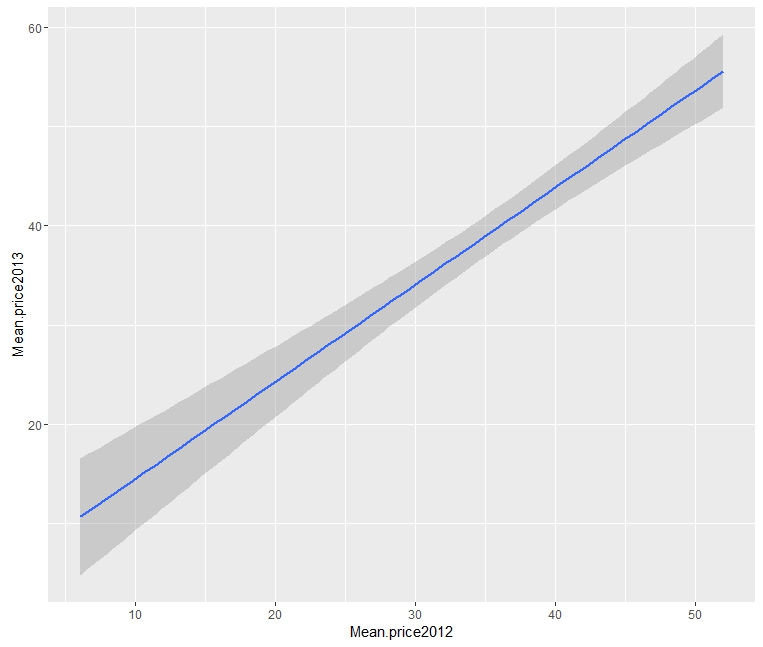
An regression analysis of the dataset was done .

* + 1. The 2 years pricing pattern for the Wholesale market results shows that there is upward positive trend of increase .



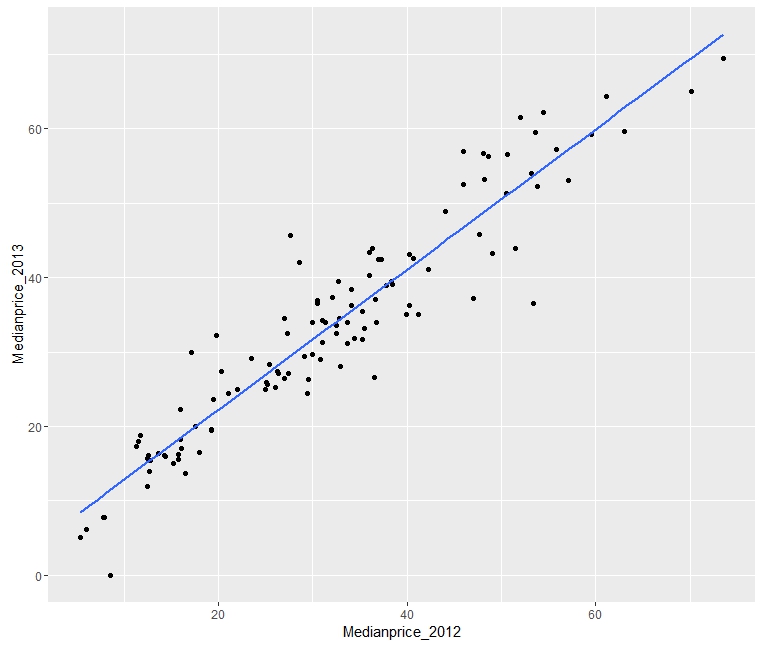
b)Organic price pattern in Atlanta 

c)Organic price pattern in San Francisco



2)The Linear regression model was applied on month-wise for two consecutive years. Median prices were used for linear regression model:

They show a positive outlook -upward price increase for produce.



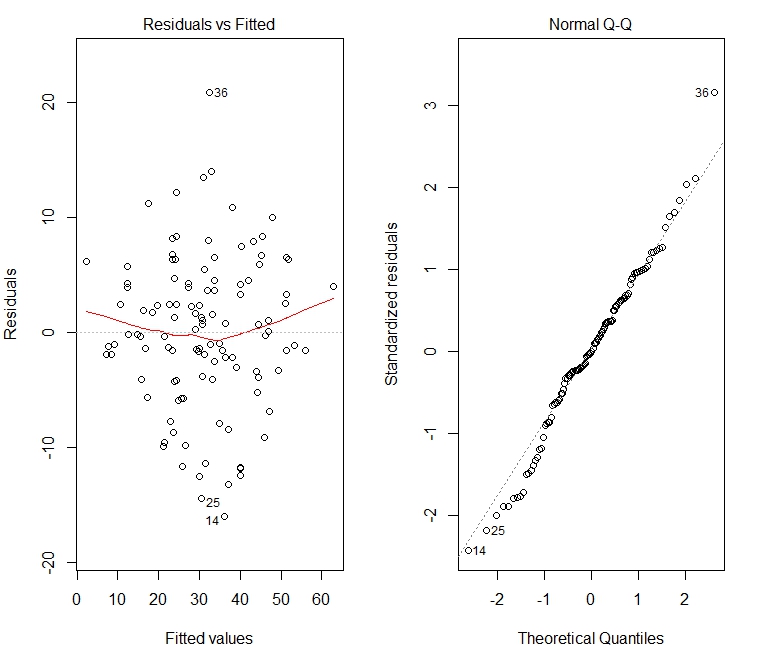
Linear regression of January Month (2012-2013)

lm(formula = January\_2012 ~ January\_2013, data = veg)

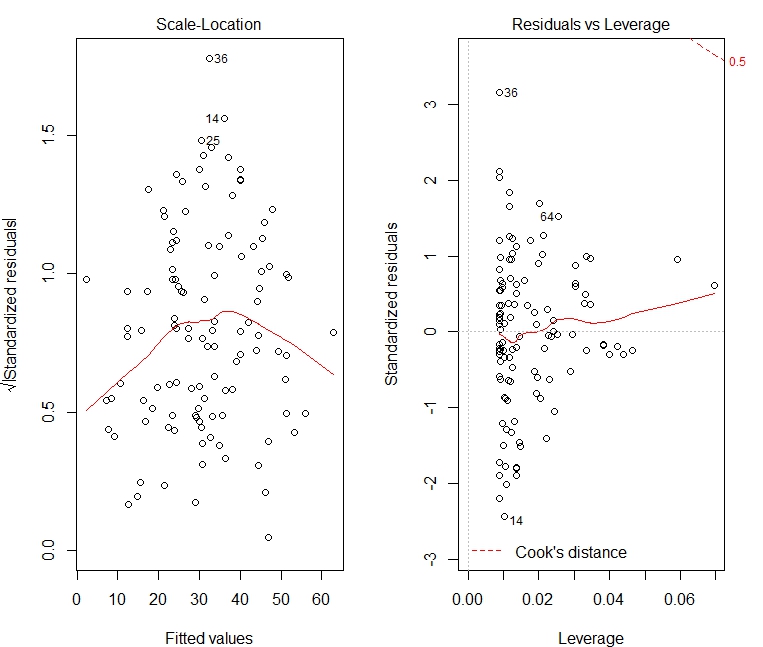
Residuals:

Min 1Q Median 3Q Max

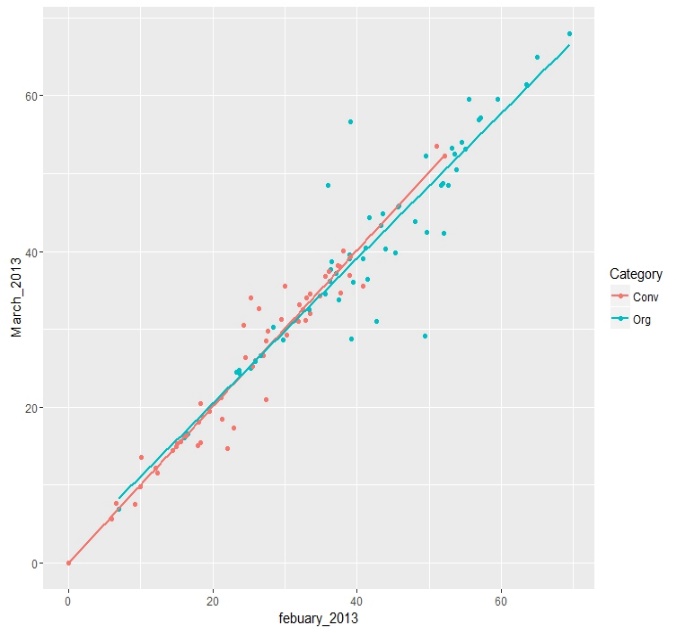
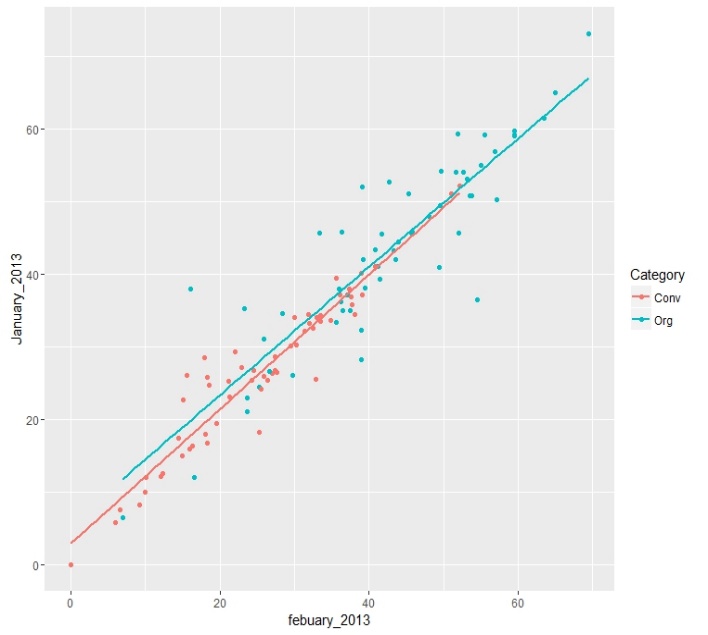
-16.0412 -3.7531 -0.0845 4.1912 20.8692

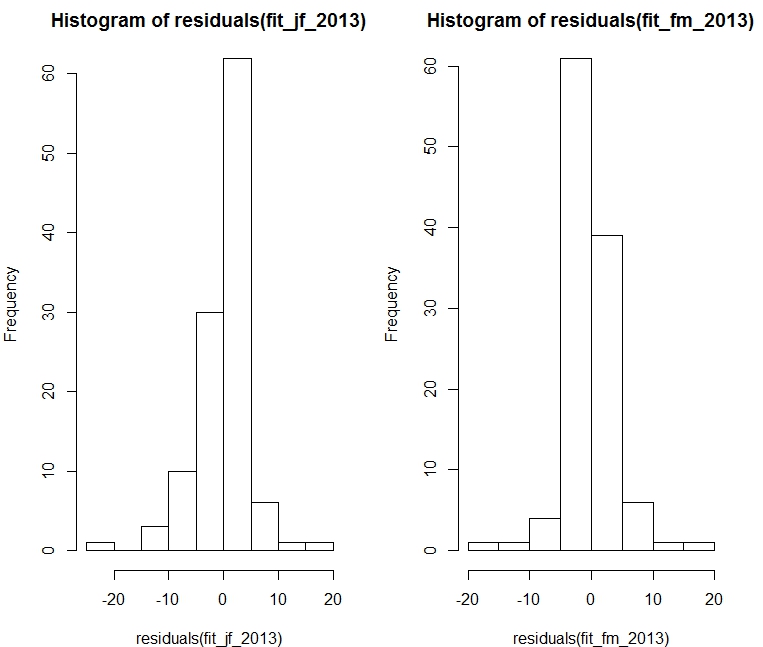


The residuals shows that the slope(B1) is significantly is below zero

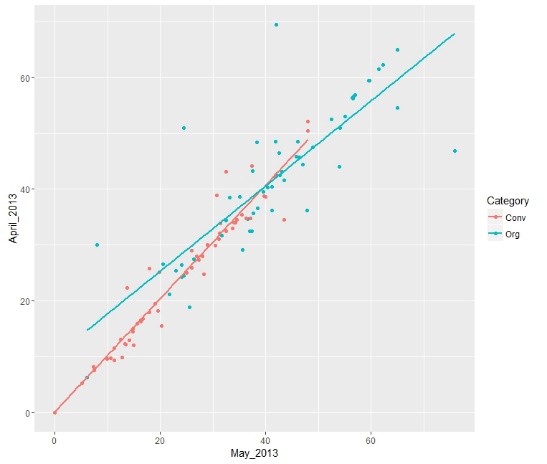


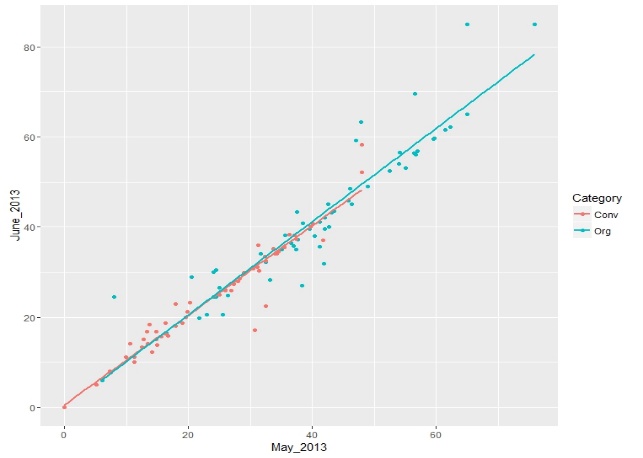
**Anova Modeling :**

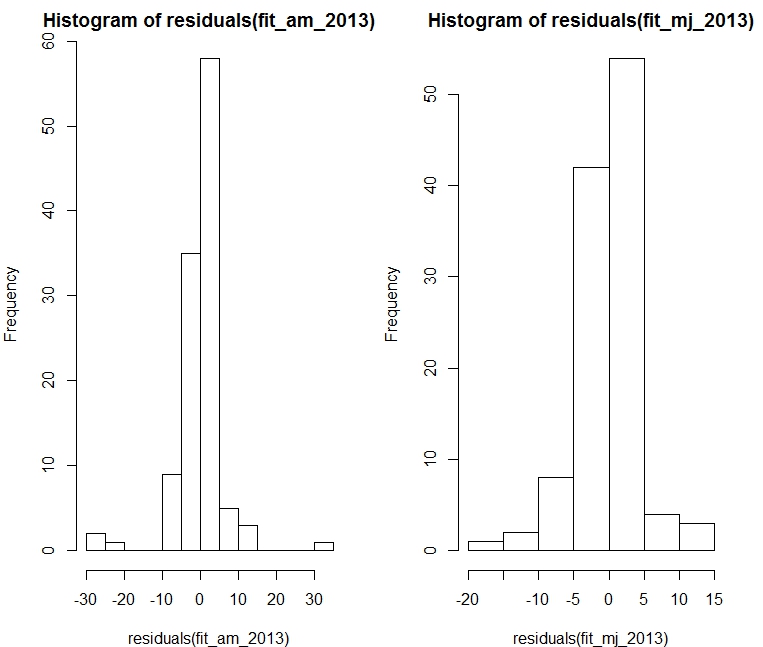
Anova test were conducted on the dataset based on each quarter of 2013. This will help in determining the Month-On-Month (M-o-M) pricing of different produce in two categories (Organic and conventional) Q1: January-February-March 2013: 



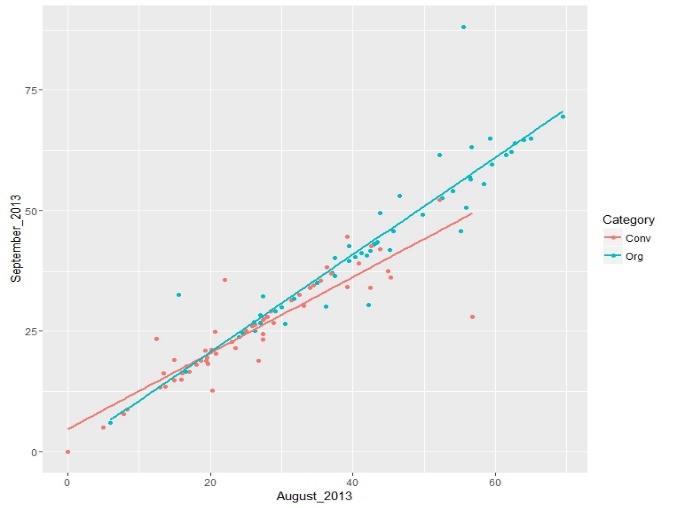
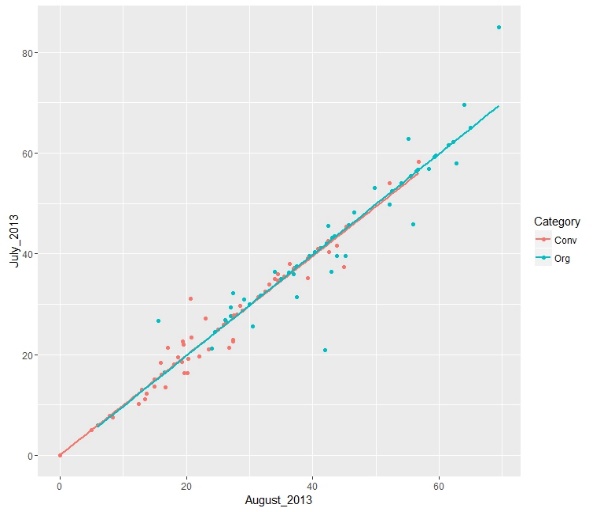
Q2: April-May-June 2013

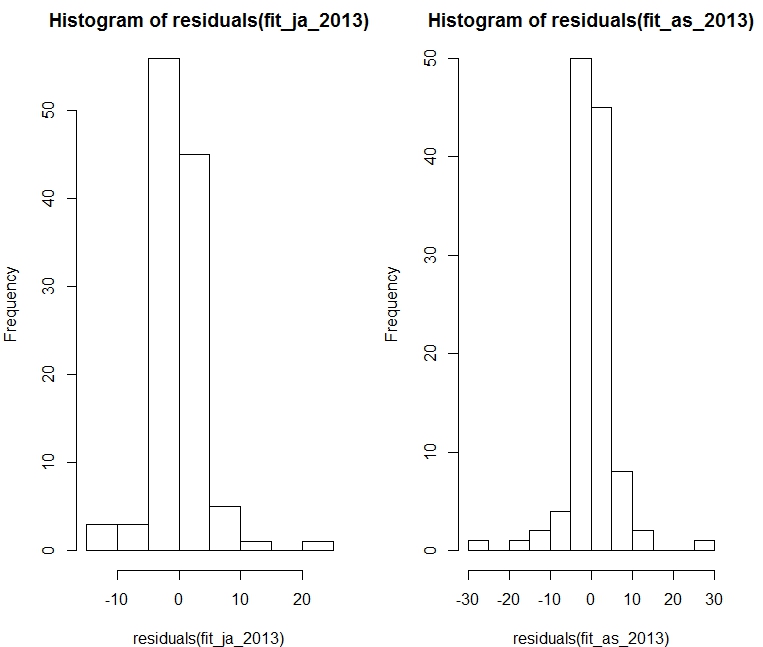




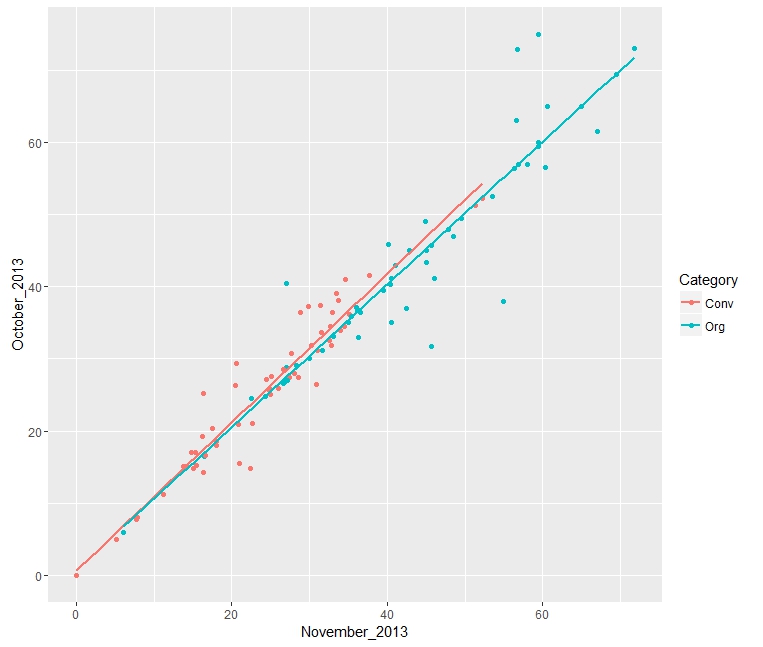


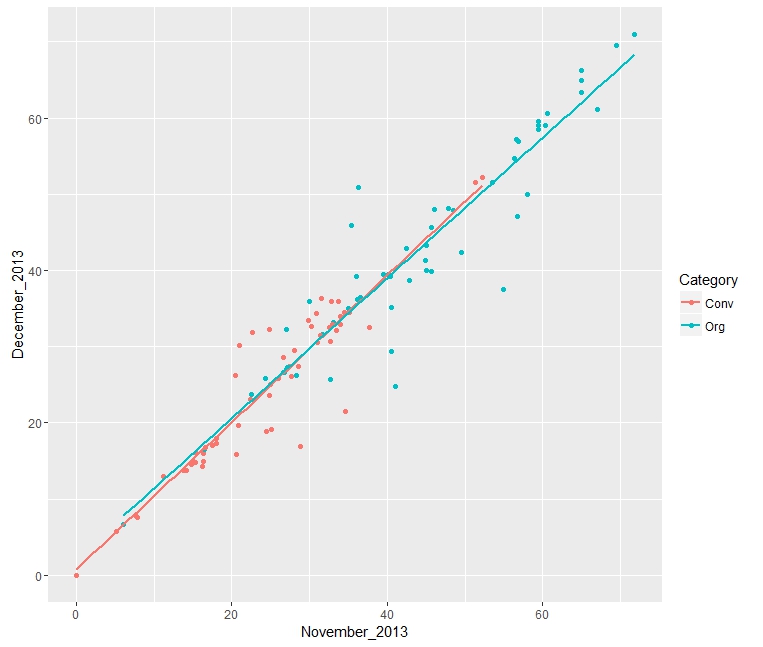
Q3:July-August-September 2013

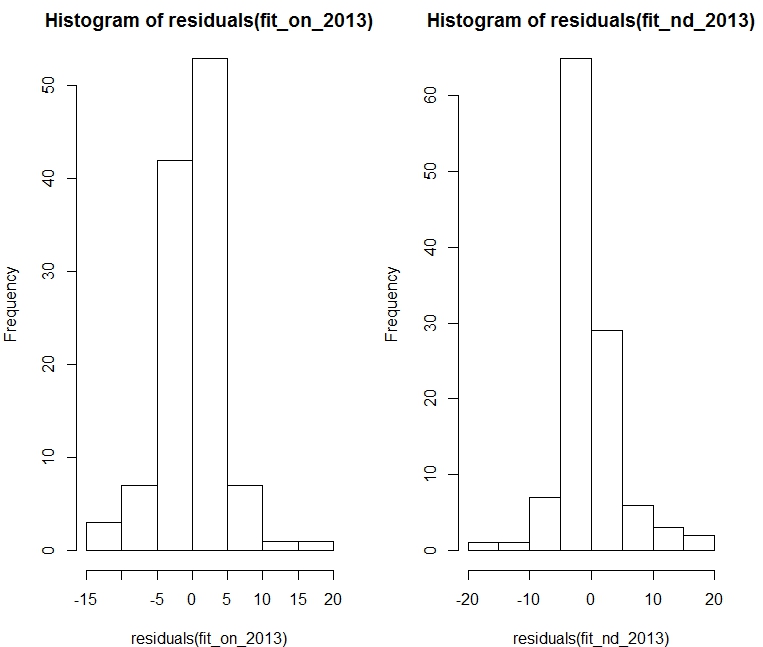




Q4:October-November-December 2013



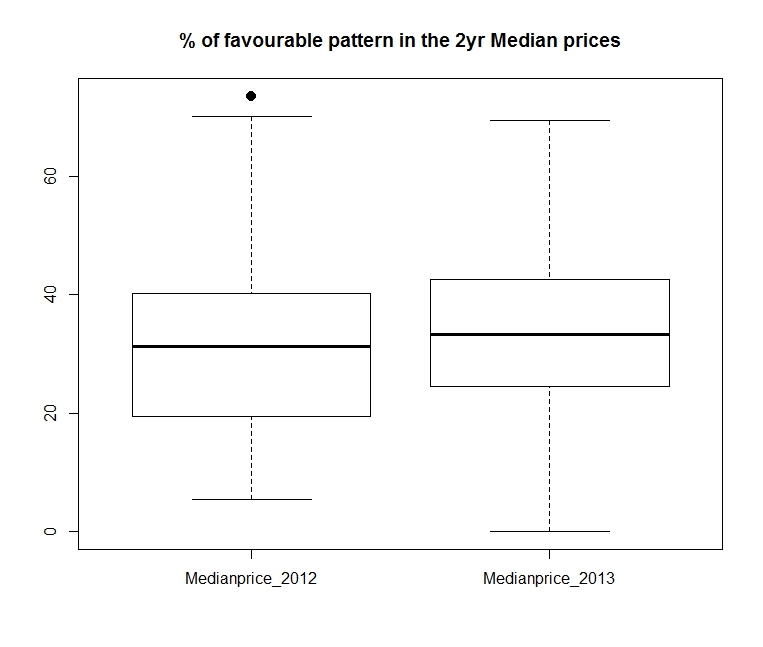




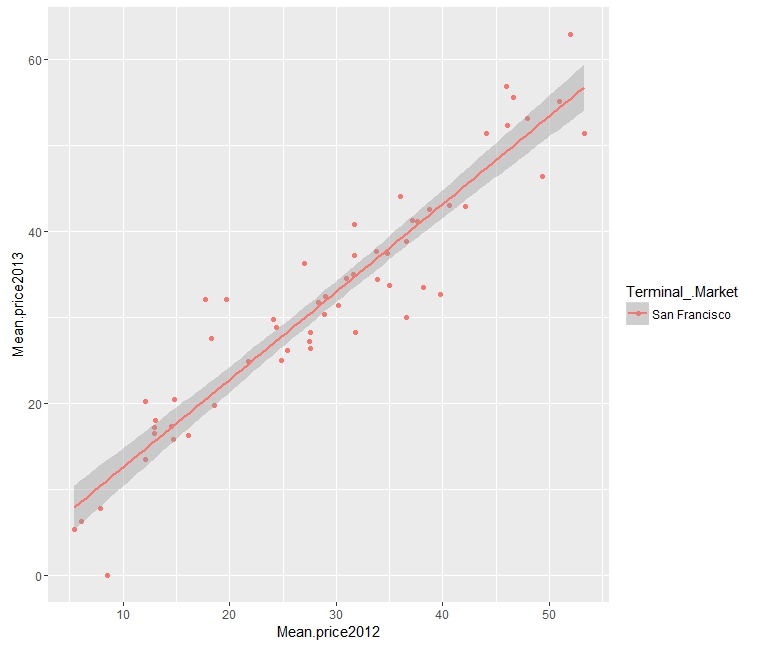
All the Anova models show a Positive strength and strong linear regression depicting the relationship between the two consecutive months are at positive linear pricing.

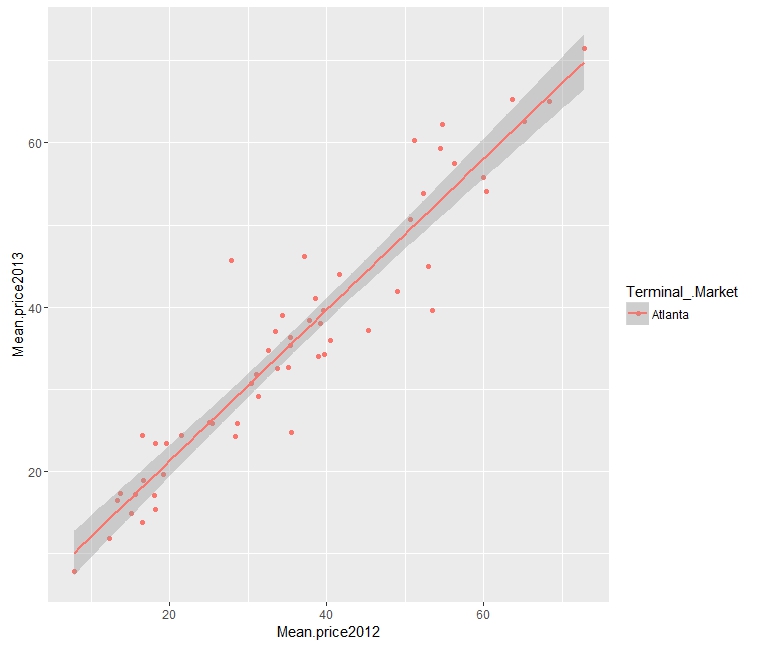
**II. K Mean clustering:**

The variables of the pricing were clustered into three different groups and the Median price pattern for two years were derived. Median being around 30 and 32 respectively.

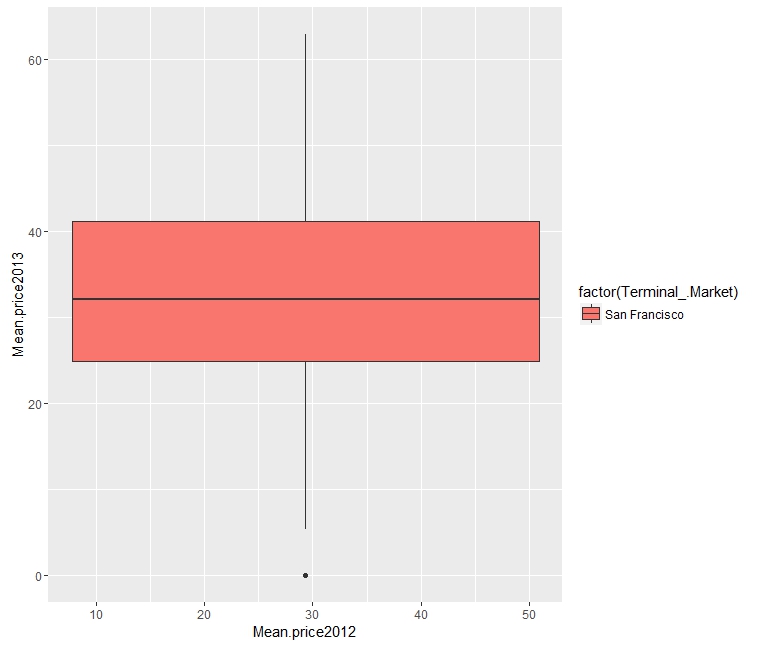


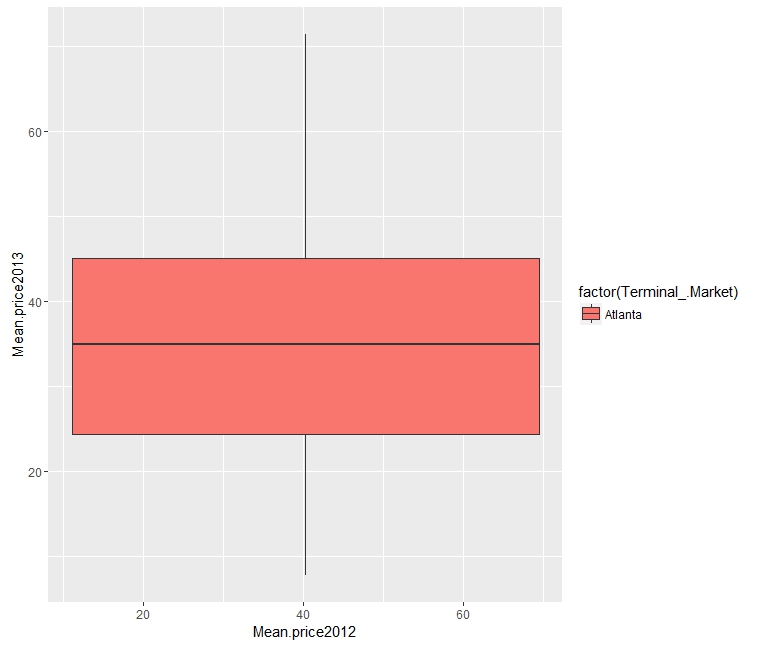
**III. Location Analysis (The pricing pattern between West coast and east coast)**





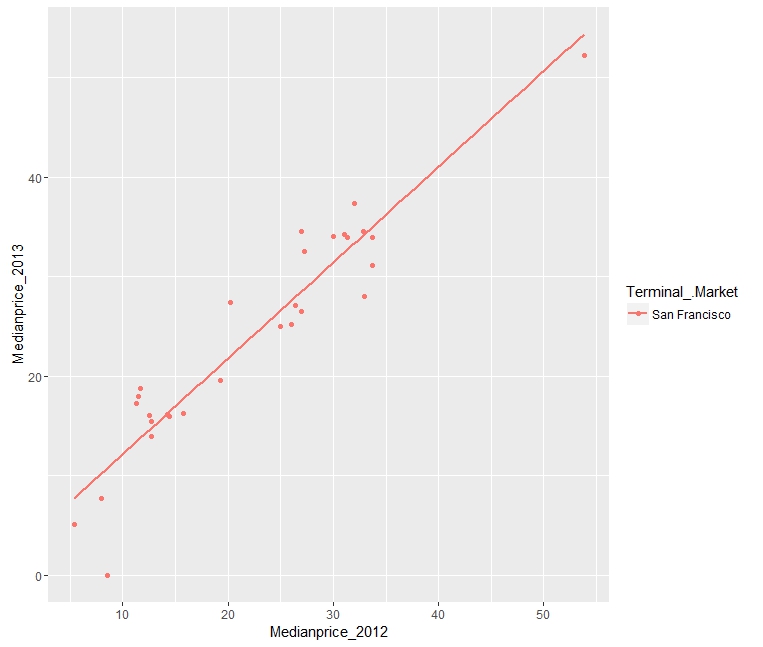
Boxplots of San Francisco and Atlanta depicts the consistency in the price pattern between both the geographical location.

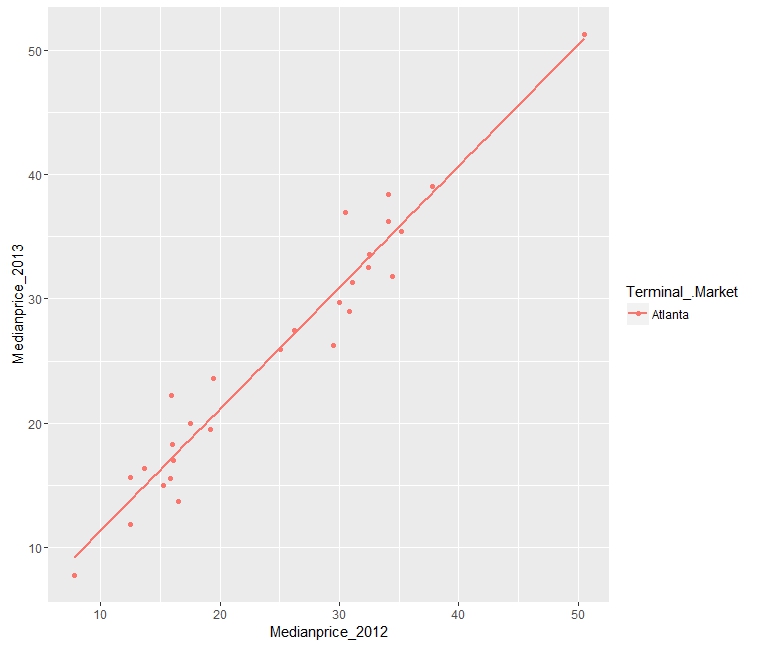




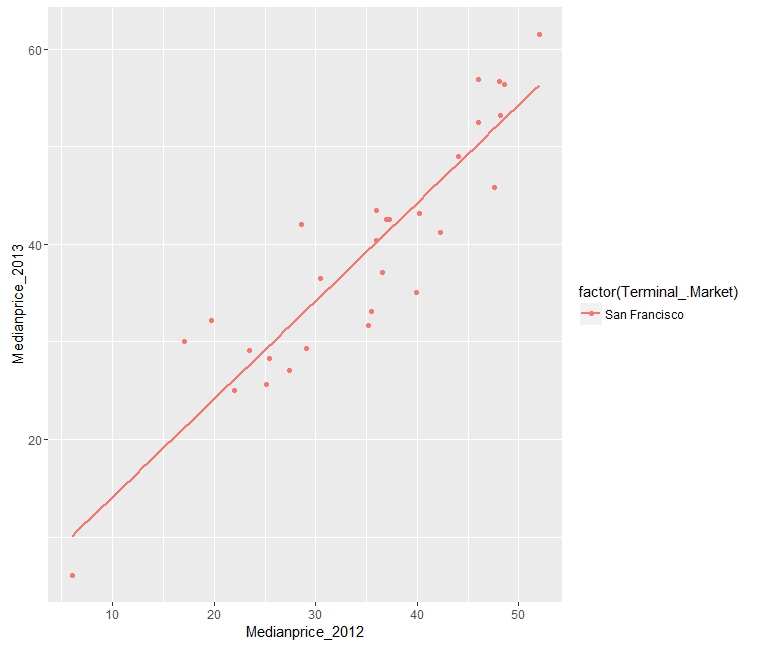
**IV. Organic and Conventional price differentiation**

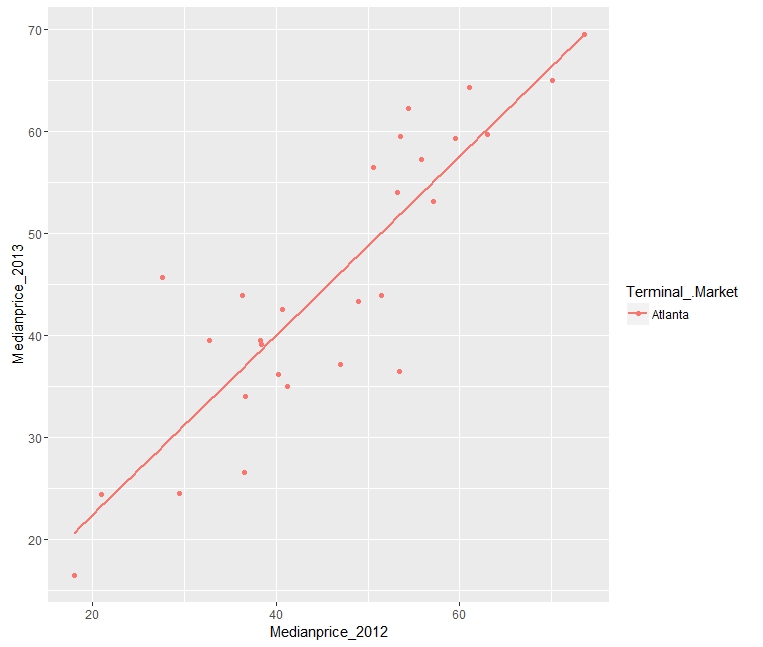
Conventional price pattern in SFO and ATL:





Organic price pattern in SFO and ATL:

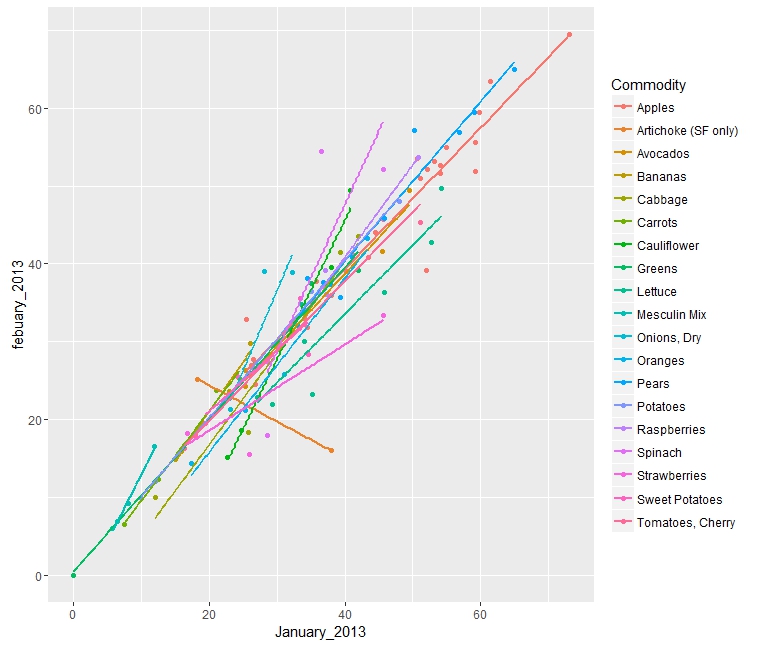




There is no significant Price differentiation in Conventional and organic pricing in both cities.

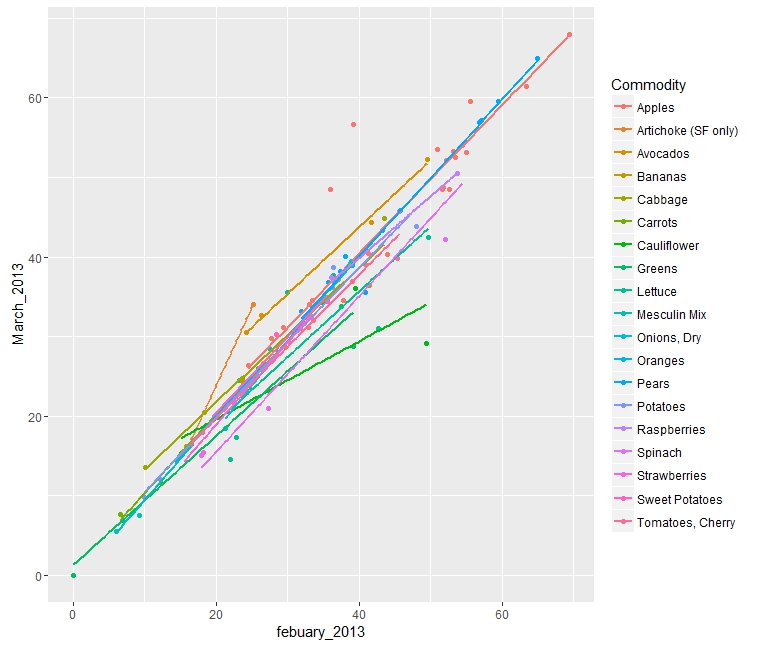
V. Month-on-Month (Commodity-wise) Pricing pattern

January 2013- February 2013

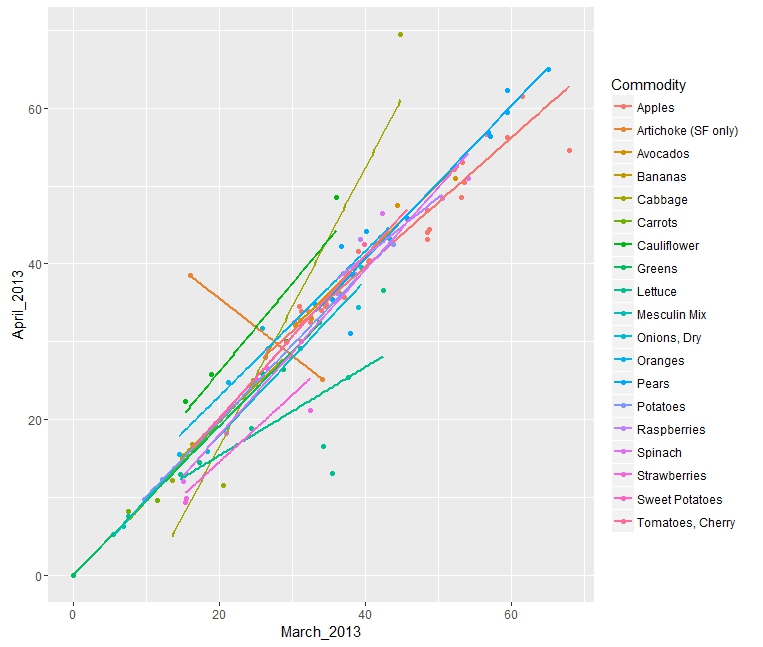


Expensive: Apples and Oranges

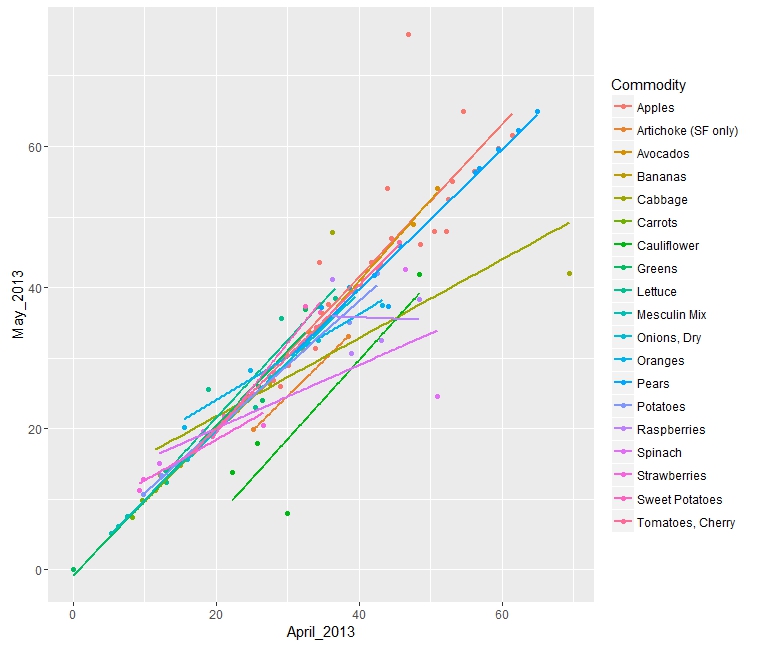
Inexpensive: lettuce and Greens

February and March 2013 

March and April 2013



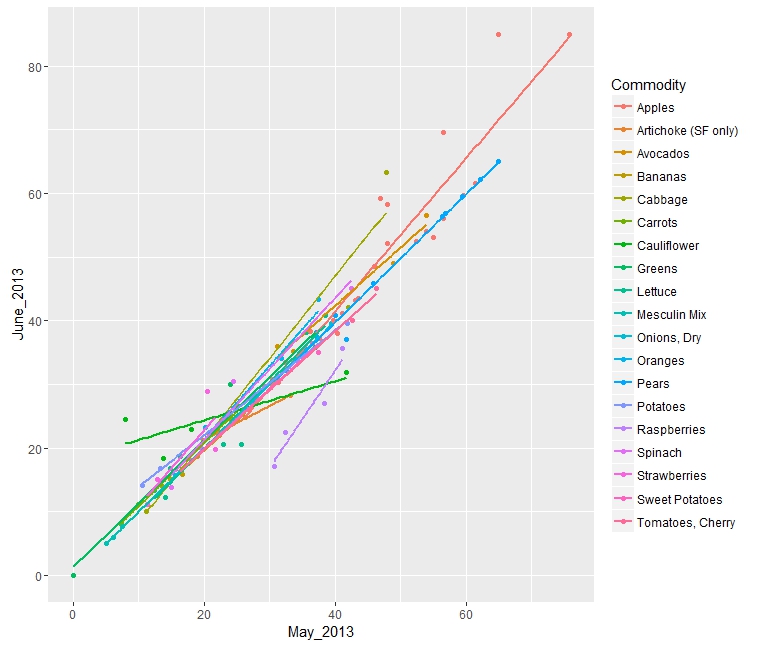
April and May 2013



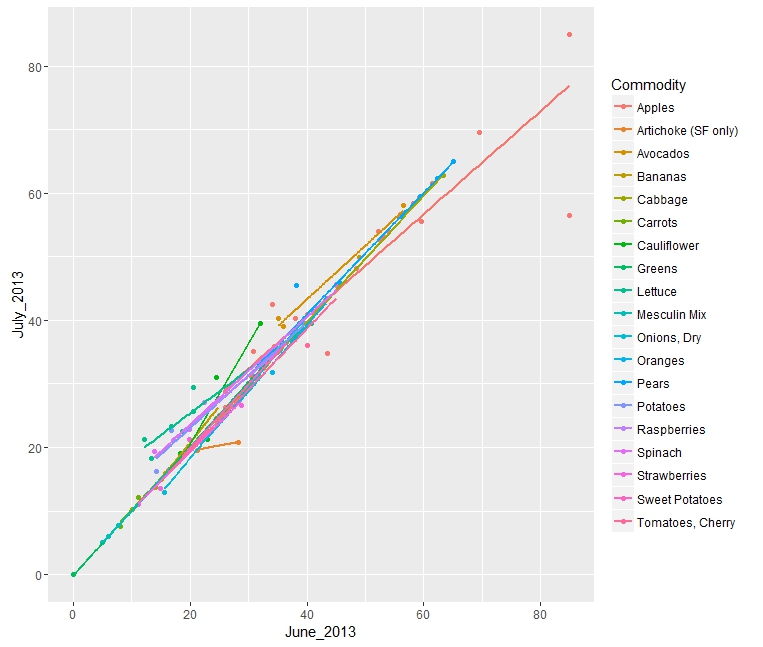
Expensive: Apples and Oranges

Inexpensive:Cauliflower

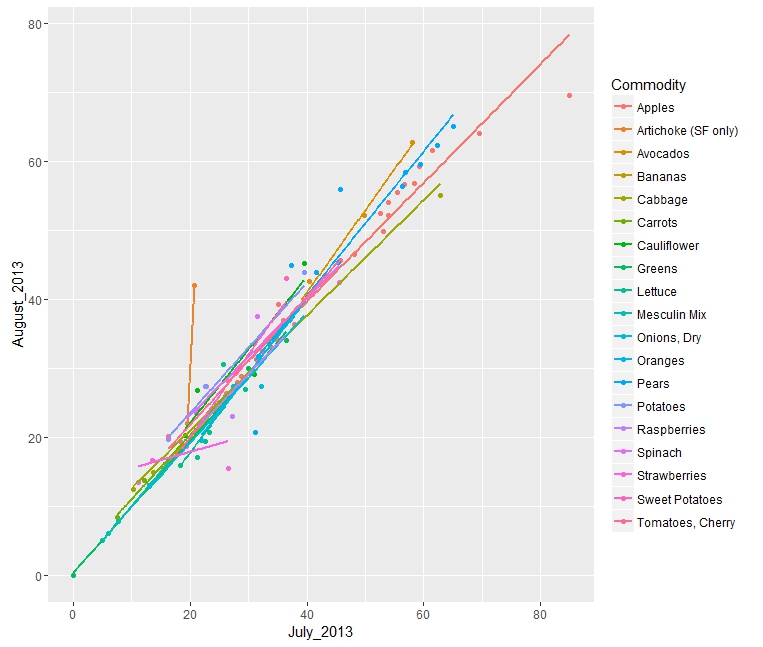
May and June 2013:



June and July 2013:



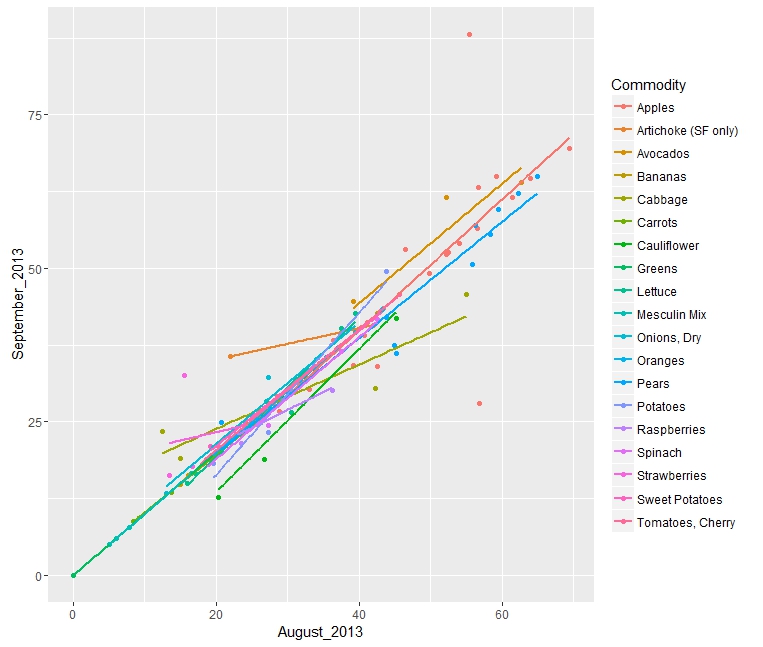
July and August 2013:



Expensive: Apples Inexpensive: Greens

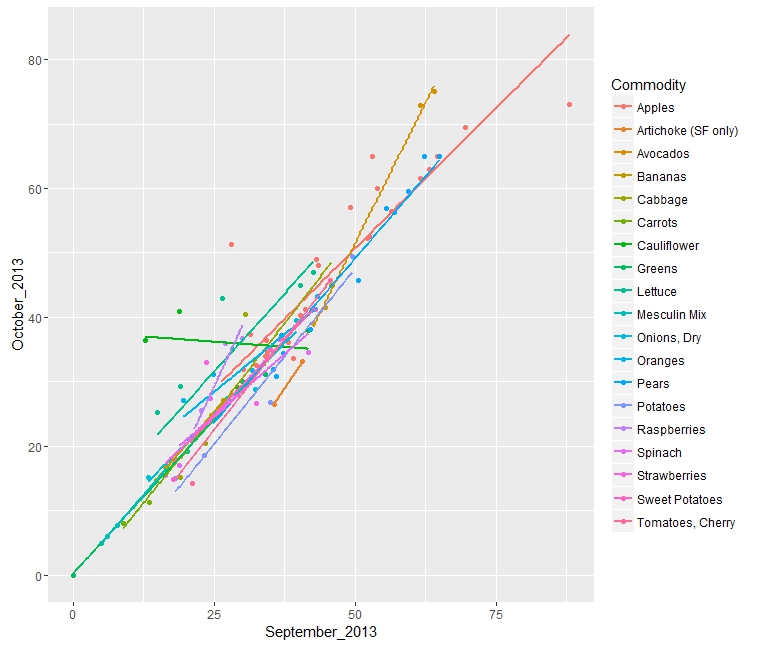
Orange prices have reduced compared to Apples

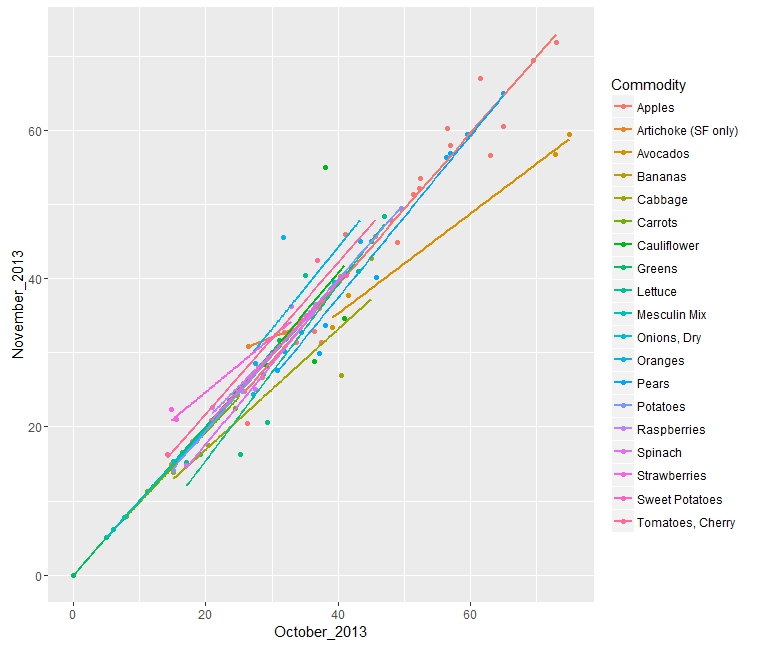
August and September 2013:



Apples are still leading in the pricing pattern. Bananas has a higher pricing

September and October 2013:



October\_November 2013:

November and December 2013:



Overall, the pricing pattern depicts the trend in the produce:

Apples being the expensive produce and Greens to be least expensive.

# Evaluation and Deployment

Outcome of Statistical Analysis:

1)Pattern in produce pricing – A comparison of Organic and conventional Produce

**Pricing Pattern is comparatively the same and consistent, there is no significant price difference between organic and conventional.**

2)Do the Geo-graphical locations differ or correlate in their pricing pattern?

(San Francisco and Atlanta)

**The pricing pattern is similar relatively and correlate for both geographical locations.**

3)(consumers can be selective on the Month-wise pricing) example: Restaurant owners can decide which produce to buy according to the Price increase-decrease and their operating cost.

**Yes, Month-wise price pattern can determine on consumers being selective on purchase**

|  |  |  |  |
| --- | --- | --- | --- |
| Month | Expensive | moderate | Inexpensive |
| January | Apples | Raspberries | Greens |
| February | Apples | Strawberries | Cauliflower |
| March | Apples | Onions | Greens |
| April | Oranges | Spinach | Greens |
| May | Apples | Tomatoes | Lettuce |
| June | Apples | Potatoes | Greens |
| July | Apples | Strawberries | Greens |
| August | Apples | Cabbage | Lettuce |
| September | Apples | Onion | Greens |
| October | Apples | Carrots | Greens |
| November | Apples | Cauliflower | Greens |
| December | Apples | Strawberries | Greens |

4)Predictive analysis of prices will help the consumer on the produce price that are probable to increase

**Yes, Predictive analysis shows that both organic and conventional prices have linear positive upward pattern. Therefore, increase in prices is the trend.**

5)Overall perspective on Organic produce.

a) Organic and conventional have the same pricing pattern.

b) This result gives a different perception of Organic prices being expensive than conventional prices.

**Price variation will not affect the decision of purchasing organic produce, bearing in mind the change of global switch-over or outlook to Organic produce**.

It is suggested that the wholesale traders need to delegate procurement cost to purchase Organic produce instead of conventional produce.

This is the final Inference. In addition, Organic produce has its benefits:

Why Organic?

1. To avoid exposure to chemical pesticide residues
2. Help support local farming and available produce
3. Nutritional benefits
4. 8/10 individual consumers buy organic

Why is Organic being perceived to be expensive?

1)Post-harvest logistics (usually small batches) propagates increase in distribution cost

2)Distribution chain is dis-organized

3)production cost is higher

Why is conventional produce perceived to be less expensive?

1)Environmental fertility is not considered

2)In-appropriate handling of pesticides.

# 10.References

1)<https://www.carolinafarmstewards.org/wpcontent/uploads/2015/05/Wholesale_Packing_Resource_Guide-1.pdf>

2)<http://www.consumerreports.org/cro/news/2015/03/cost-of-organic-food/index.htm>

3)<http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=2455&context=etd>

4)<http://www.choicesmagazine.org/2007-2/grabbag/2007-2-05.htm>

5)<https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/>

6)<http://www.fao.org/organicag/oa-faq/oa-faq5/en/>

7)<http://ageconsearch.umn.edu/bitstream/56659/2/jaae413a01.pdf>

8)<http://www.foxnews.com/leisure/2012/03/11/10-reasons-organic-food-is-so-expensive/>

9)<http://www.ecowatch.com/10-reasons-consumers-buy-organic-1881899943.html>

10)<http://www.terminalmarkets.com/markets.htm>

11)<http://www.sfproduce.org/merchants/merchname.html>

12)<http://terminalmarkets.com/atlantastate.htm>

13)<http://www.ers.usda.gov/media/255736/aib777c_1_.pdf>

14) Data Science for Business – Foster Provost and Tom Fawcett