Analysis of tomato prices in Karnataka

INTRODUCTION

The horticulture market in India is one of the crucial, yet understandably most unpredictable fields. Everyday tonnes of goods are shipped to and from the markets, which play a key role in connecting the consumers and the producers. They also play a role in deciding the prices, because unlike the markets for cereals, which have a minimum government rate, the horticulture market, consisting of easily perishable goods has its prices fixed on a daily basis, depending on the demand and supply.

Historically, agriculture was done by observing nature, but with time that knowledge is slowly degrading and many farmers are looking for other occupations. Climate changes are causing unpredictable rains, this adversely affects yields, which affects market prices. Hence it is necessary to understand the changes that are happening, to take cautionary measures in the near future, to ensure food supply.

The dataset was obtained from Kaggle (link in reference). It is about the prices of tomatoes in Karnataka (2016-2018). The dataset consists of 11421 rows and 7 columns which are as follows:

1) Market Area – 40 areas in Karnataka
2) Date
3) Tonnes
4) Variety – 2 types, tomato and hybrid
5) Minimum Price (Rs. /Quintal)
6) Maximum Price (Rs. /Quintal)
7) Modal Price (Rs. /Quintal)

UNDERSTANDING DATA DISTRIBUTION

First, we group the data, month wise by taking the mean of the tonnes and modal prices, to get an idea of the monthly average prices and quantity. It is observed that some of the rows have null, values. These values are dropped of for the analysis, the number of rows is now 11286. The distribution plot for prices, tonnes, variety and markets are plotted below. The bars represent bins of histogram, and line represents kernel density estimation (KDE). The y axis is density and x axis show values.

INTERPRETATION OF PLOTS:

1) It can be observed that prices are mostly around Rs.500 to 700. Then Average price per month per quintal is Rs 600
2) Mostly there seems to be around 50 tonnes inflow to markets

Fig1: distribution of prices

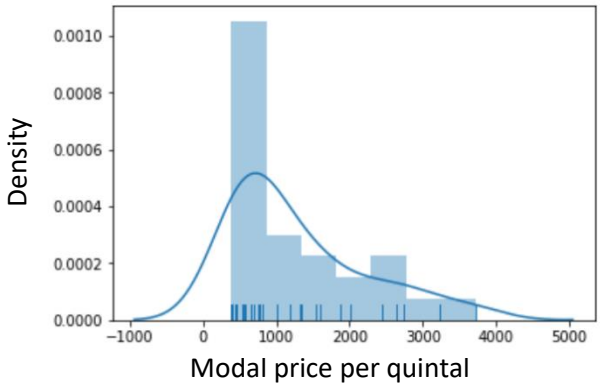
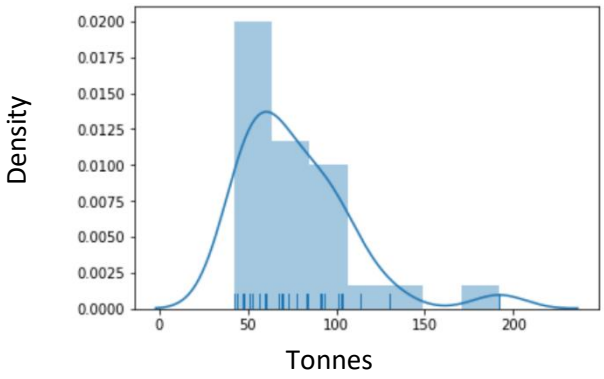


Fig2: distribution of tonnes



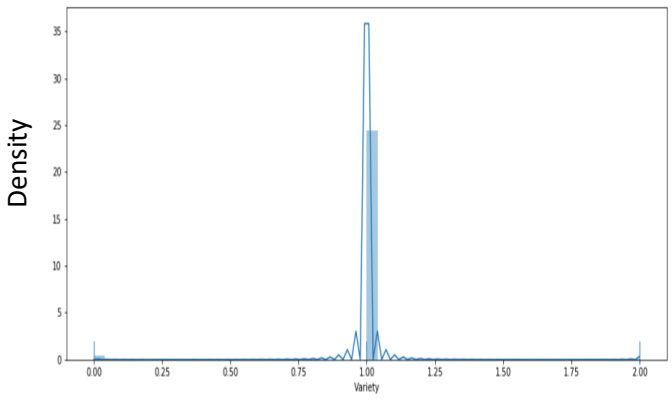


Fig3: Distribution of variety

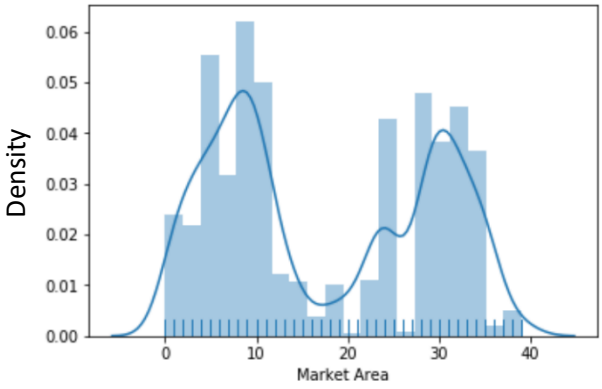


Fig4: Distribution of markets

3) The data with respect to variety is highly skewed with very little data on Hybrid variety to make significant conclusions, hence it is neglected
4) Of the 40 markets some markets have very less data, hence markets with less than 80 values are neglected

VISUALIZING DATA

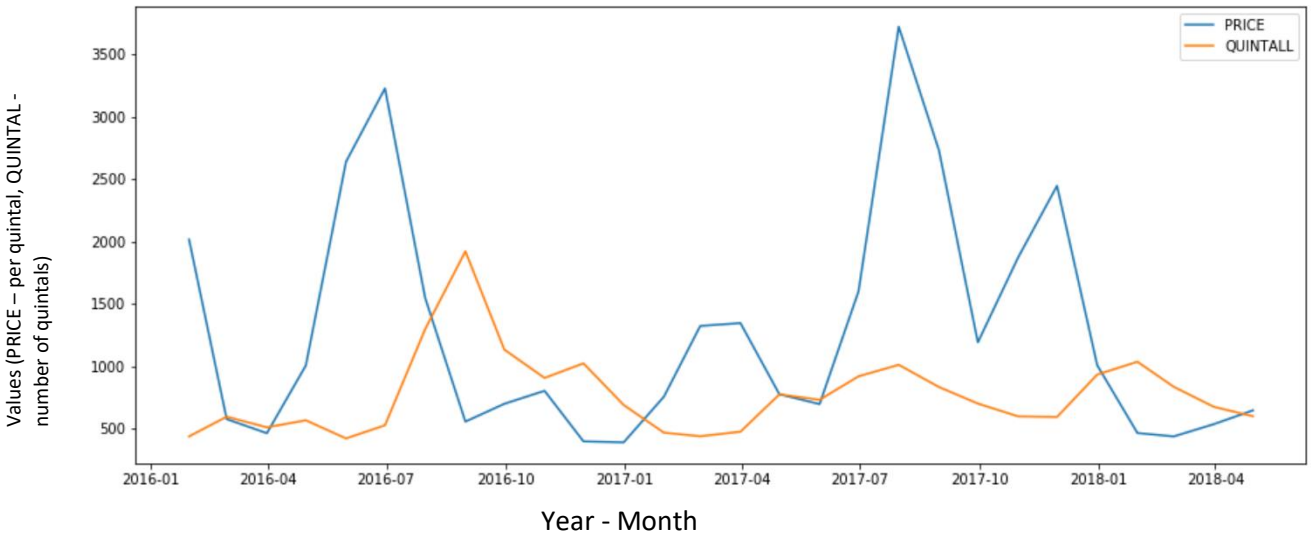


Fig 5: Plot of Tonnes and Prices monthly (in quintals)

Interpretation:

On plotting the average Tonnes and Prices per month we observe the following:

1) We can clearly see the rise in prices with fall in supply and vice versa
2) In July 2016 there is sharp rise in prices, and a similar peak is observed in August 2017.
3) The rise in supply in September 2016 is matched with a steep fall in prices. While in 2017, the supply was maintained below 100 tonnes which has led the prices to be higher than 2016.
4) Surprisingly, in 2016, the beginning of the year had high prices, which steadily came down until April.
5) In 2017 the prices steadily increased from January, until April.
6) Again in 2018, the prices slowly rise from January to April.

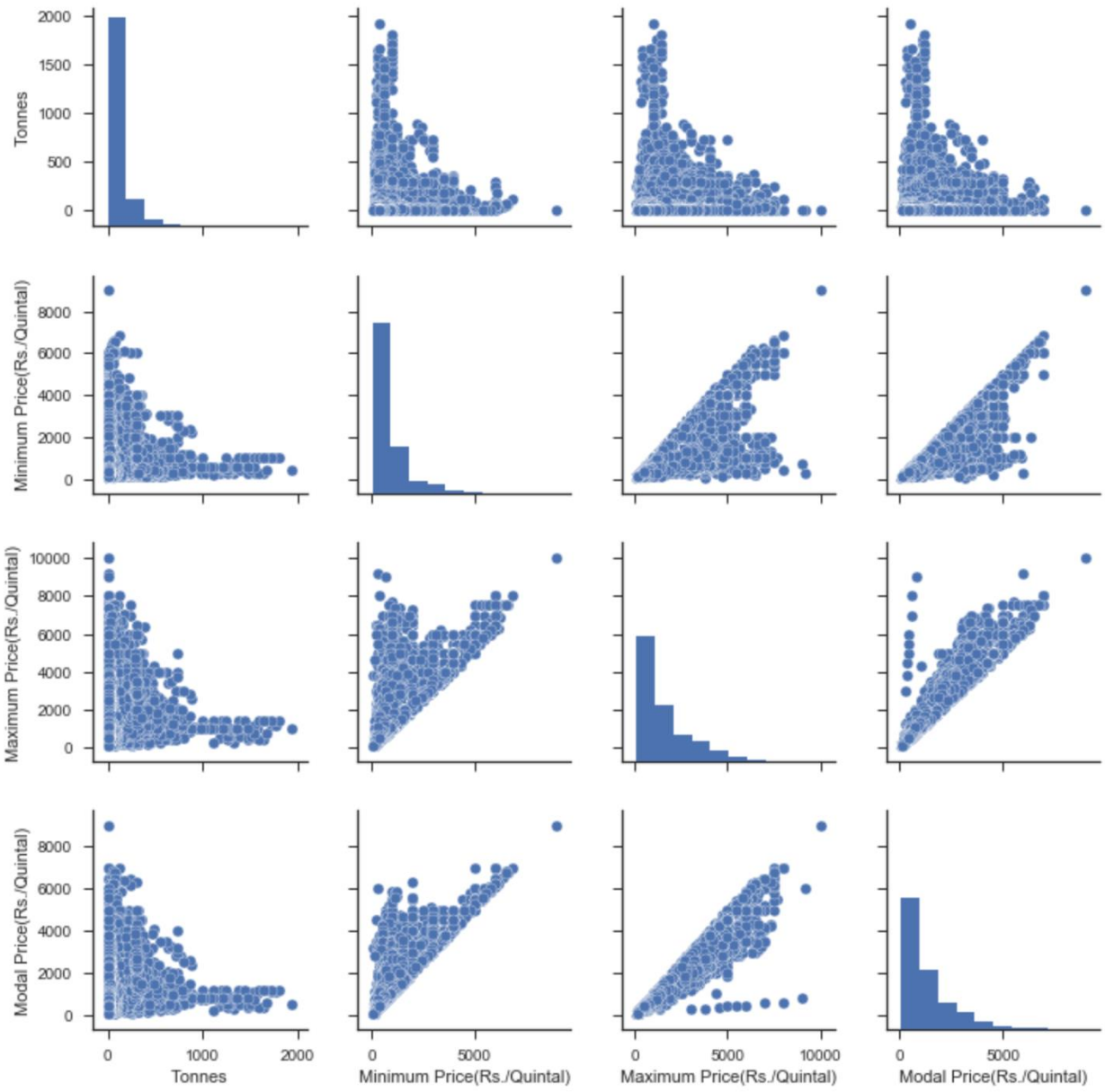


Fig: 6 Visualizing relationships

Interpretation:

We find that the graphs of Tonnes vs Prices (all three types) are similar. The graph is in the shape of an exponential decay graph.

Minimum prices are mostly below 5000, while maximum prices have many data points above 5000.

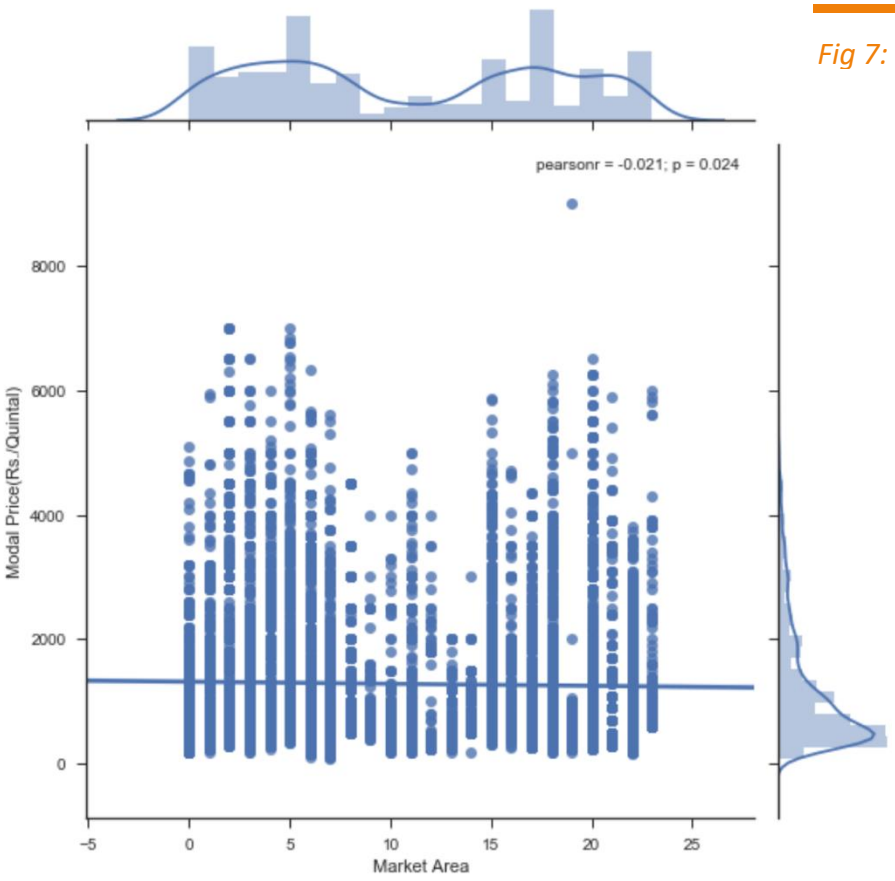


Fig 7: Prices vs Market area

The graph shows marketwise prices of tomatoes. It also shows the distribution plot and KDE of markets on top and distribution plot and KDE of prices on right side of graph

Interpretation:

The average price among all markets is approximately Rs 1500 per quintal.

The depression in prices for markets 7 – 10, 11-14 and 19 can be due to the lack of data or size of markets.

Fig 8: Tonnes vs Market Area

The graph shows the tonnes received market wise. The line is a regression line to see how the data fits. On top is a distribution plot with KDE for market areas, and on right is a distribution plot with kde for tonnes

Interpretation:

It is very evident that market area 15, which corresponds to Kolar market, receives the highest number of tonnes. Also, most of the markets (13/25 * 100) 52% markets receive less than 60 tonnes of tomatoes

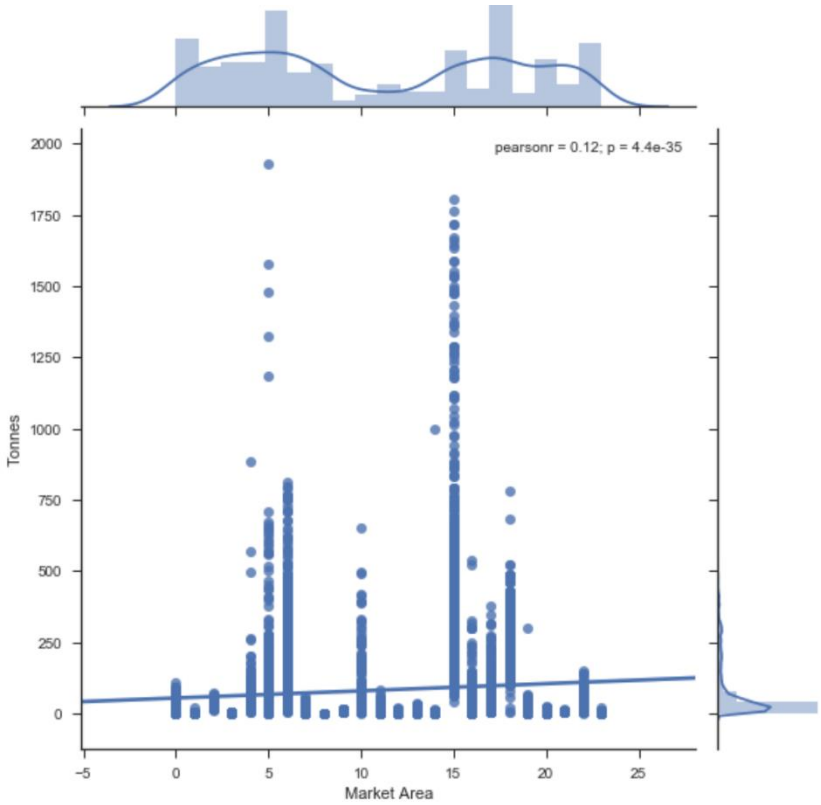
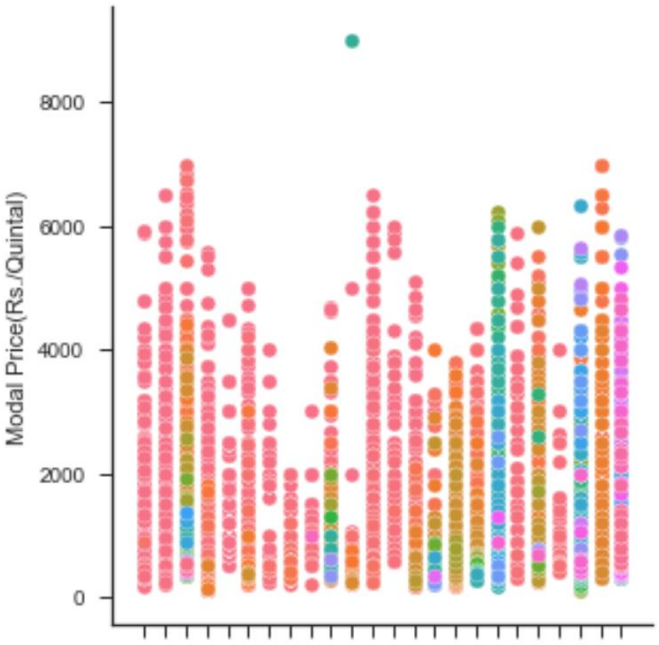


Fig 9: Prices vs market area (with colour representing tonnes)



Interpretation:

Here the prices are plotted marketwise, with color representing the tonnes received. It is noticed that most markets receive 1 – 15 tonnes, which is indicated in red colour. Also, the prices for the same quantity is different for different days in different markets.

In market 3 the higher tonnes face lower prices, while lower tonnes (10-20) face the highest prices.

In market 19, higher values of tonnes is received, and has higher prices, when compared to other markets receiving similar tonnes.

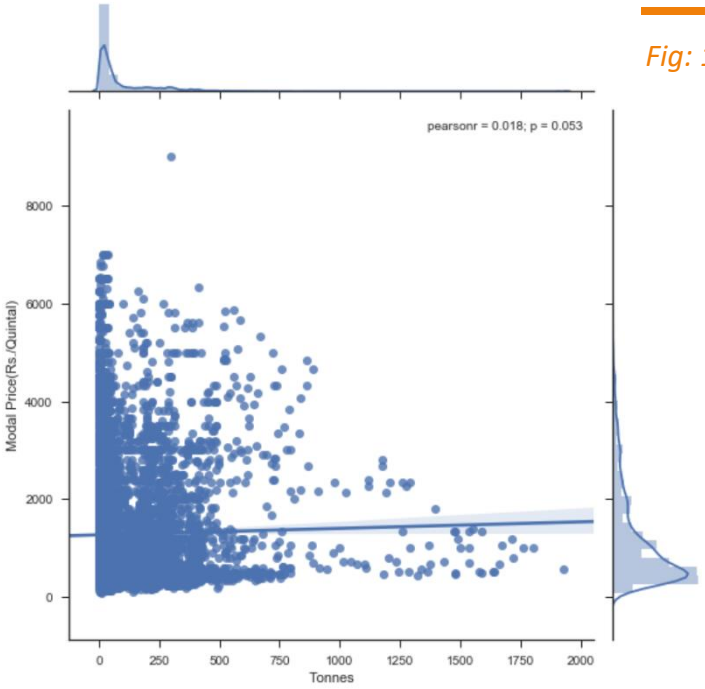


Fig: 10 Prices vs Tonnes

The graph shows relationship between prices of tomatoes and Tonnes. It also shows the distribution plot and KDE of Tonnes on top and distribution plot and KDE of prices on right side of graph. The line is a regression line, to see how the data fits.

Interpretation:

It is very clear that, for very low tonnes, prices are very, high and a small change of tonnes (about 10) brings down prices by thousands. The tonnes mostly do not exceed 750. Highest tonnes (more than 1000) face very low prices.

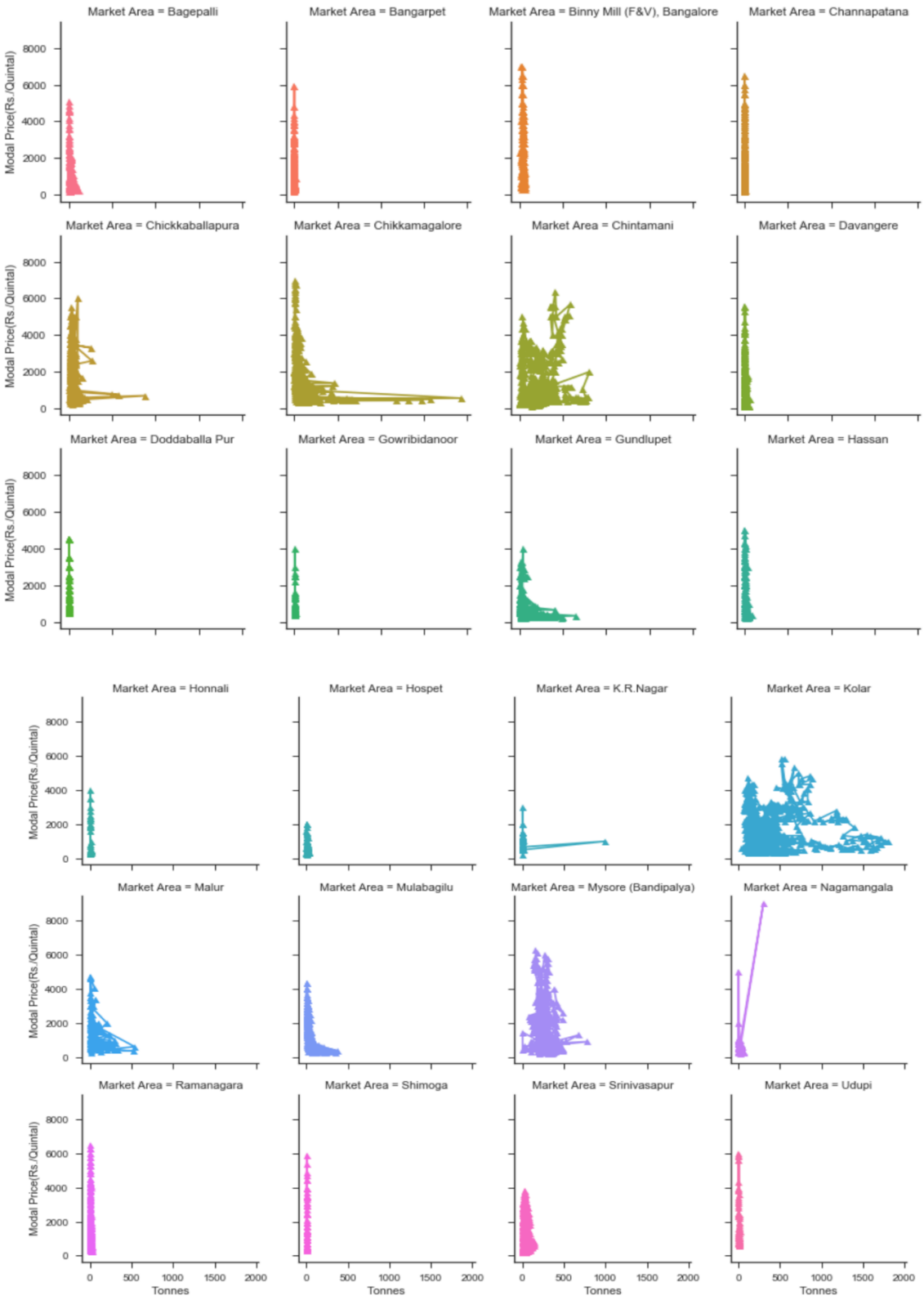


Fig 11: Area wise visualization of prices and Tonnes

Interpretation:

This graph shows the marketwise relationship between Price and Tonnes. Most markets receive less than 100 tonnes.

Doddabalapur, Gundupet, Malur, Shrinivasapur markets have received up to 500 tonnes. Chikkamagalore market has few instances of up to 2000 tonnes.

The largest receiving market is Kolar.

It is very evident that the same quantity receives different rates on different days, across all markets. It is observed that Maximum, minimum, and modal prices are highly (80%) correlated.

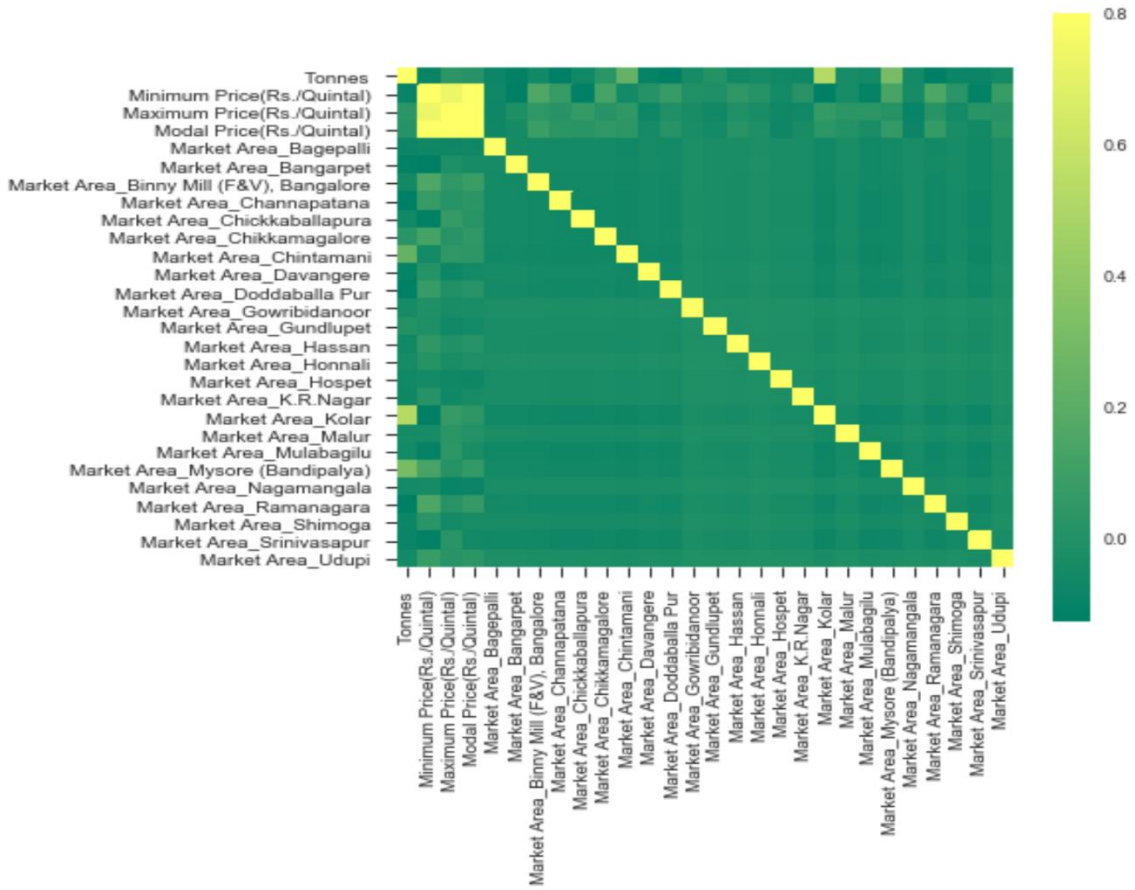


Fig 12: Correlation map

Interpretation:

Kolar market and tonnes is also correlated (60%). This shows that by knowing any one type of price, we can easily estimate the others, Higher tonnes are mostly received by Kolar market, as it is positively correlated to Tonne

ANALYSIS

The data that was used has been modified by removing rows which had null values and removing markets that had too little data points. This resulted in a dataset containing 11286 rows, and 25 market areas.

The trend between price and tonnes shows that as tonnes increase, prices decrease but this is not followed throughout as during the month of July prices seem to increase irrespective of tonnes for both 2016 While a similar phenomenon is observed in August in 2017.

The markets mostly receive 60 tonnes. Prices seem to reduce when the quantity increases above this, except for July in 2016 and August in 2017.

The average price among all markets is found to be Rs 600 per quintal. The approximate value of tomatoes per month can be calculated as follows: 600 x 60 x 10 = 360000. Hence the monthly value is about 3.5 lakhs, while yearly value is approximately 42 lakhs. This is only when tonnes are maintained below 60.

Kolar markets is shown to be the largest market with highest receiving tonnes. This is in tune with the fact that Kolar market is the largest and main tomato markets in Karnataka.

The three types of prices: Modal, Maximum and Minimum are highly correlated to each other, and Tonnes is positively correlated to Kolar market.

CONCLUSION

Hence, we find that there is a need for about 60 tonnes per month (except for few months in between), So by maintaining the supply to this level, we can try to avoid very low prices which pushes farmers into debt. Also, it is show that monthly share translates to about 3.5 lakhs, or Rs 6000 per tonne. So, if 60 farmers send 1 tonne each to market, then they earn Rs 6000. This is further reduced if many farmers send in their supplies. At such scenarios, it would be beneficial if the farmers send their produce to other markets to avoid harsh prices.

This information can also be used to take precautionary measures and avoid sowing tomatoes at a particular season, as it is known beforehand that the market requirement can be satisfied by the existing cultivations. The sowing dates can also be changed to match the market peaks.

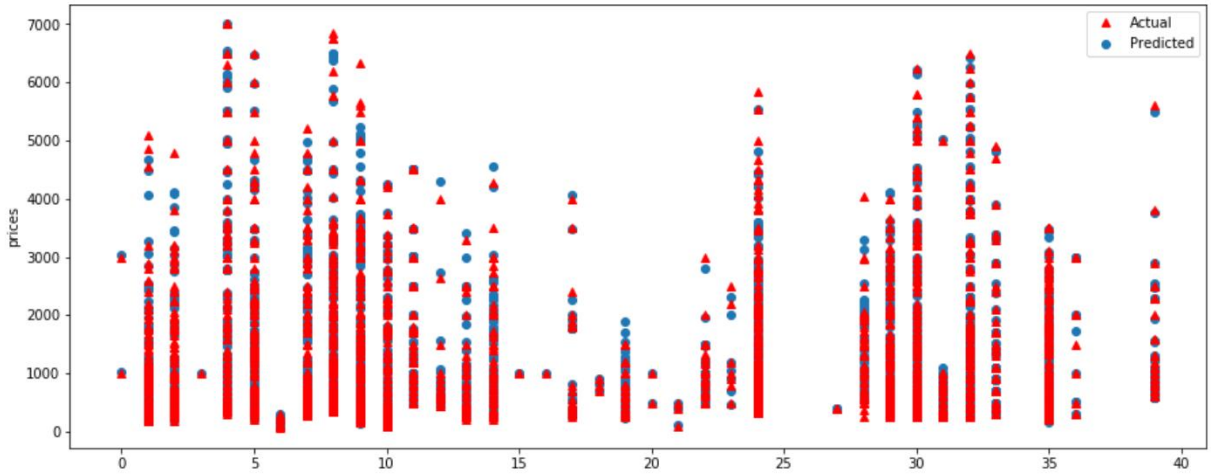
APPLYING MACHINE LEARNING

To predict the prices of tomatoes, the dataset is divided as follows:

1) X – Date, Market Area, Tonnes, Maximum and Minimum prices
2) Y – Modal prices.

Gradient boosting regression algorithm from Scikit learn in python is applied on the data and the results are as follows:

Training score = 0.9907754543010587



Test score = 0.9790274184012351

Error (square root of mean square error) = 163.39994797241687

This show that the model fits the training data 99%, and test data 98%. The predicted value may differ from the actual value by Rs 163.4

The high accuracy can be attributed to the fact that maximum and minimum prices were included in the training phase, and they are highly correlated to modal prices. This can be avoided by collecting more data with respect to weather, sowing patterns, pests, etc. More data will help in building a more robust model.

The training results upon removing maximum and minimum price feature is as follows:

- Training score = 0.8371931367310488
- Test score = 0.6523737110205012
- Error (square root of mean square error) = 319.74

Further Developments

The analysis can be further improved by including data about the agriculture fields which support these markets. The weather and sowing patterns there. Information on insects and pests that affect the crop and variety and frequency of sowing.

Reference

The data set was obtained from Kaggle