Tomato Price Forecast

ABSTRACT

The Asia Pacific is the largest producer of tomatoes accounting for over half their global production. Middle East and Africa followed the Asia Pacific as their leading producer. The United States, on the other hand, is the largest tomato processing country with a global market share of over 35%. In 2017, tomatoes represented 70% of the total processing vegetables in the United States with 22, 011.7 million pounds selected for processing. They were also the leading crops in the United States in terms of processing-vegetable farm value, being worth USD 912 million in 2017.

This report predicts the tomato prices for the coming years using the concepts of Data Analytics, using the data of the past years.The dataset consists of prices ranging from year 2005-2020. The variations in the prices in all these years, depending on factors like natural calamities, temperature, competition, weather, location, and consumer characteristics influence tomato prices.Specifically, some of the models explored could help explain the impact of vendor interaction on farmers' markets prices.

This prediction Analysis will come handy for the investors, government, farmers, and also the consumers.

INTRODUCTION

Tomatoes are amongst the foremost widely grown crops in India. India is the world's second largest tomato producer but processes 1% of its production. This impacts farmers by way of high post-harvest losses and low returns during times of market glut. Indian tomato based product manufacturers import significant quantities of tomato pulp and paste at high prices which also entails an tariff of 30%. Existing Indian paste and pulp makers are unable to work their units at optimum capacities because of an absence of fresh tomato at the specified volumes at the proper price. Further, the kinds of tomatoes currently grown in India are generally less suitable for processing because of their caliber parameters for paste and pulp production. the general results of these constraints may be a loss of import to any or all stakeholders committed to tomato production and processing and its wider impact on local and regional economic development.

Tomatoes are warm-season crops and are sensitive to frost at any growth stage, so field planting in temperate climates occurs after the threat of frost is past in the spring or transplants are planted and grown under row covers in late spring

Further, the range of processed tomato foods is additionally expanding with the introduction and demand for several able to eat meals, curries and snack products finding favor with the Indian consumer.

PROBLEM STATEMENT AND NEED OF SOLUTION

Tomatoes, being a basic need, affects the life of not only farmers, but also the government, vendors and consumers. The frequent fluctuation in prices is something that worries everyone. This indefinite nature of the prices and production, thus, has to be minimized. This project report predicts the market prices expected for the tomatoes in coming years based on the past prices and factors.

EXPLORING THE DATASET

The dataset contains 1736552 samples of tomato pricings. The number of categorical and numerical columns are 5 and 8 respectively.

The categorical columns include:

1. State
2. District
3. Market
4. Variety
5. Commodity

The numerical columns include:

1. Tons
2. Min Price
3. Max Price
4. Modal Price
5. Month
6. Year
7. Date
8. Arrival Date

Firstly, observing what all columns are there in the datasets and what are their data types.

Object: Object format denotes that variables are categorical. Categorical attributes in our dataset are mentioned above.

Int64: Represents integer variables. ApplicantIncome is the only attribute of type int64.

Float64: Represents decimal valued or numerical variables.

Datetime64 : Represents date .

COLLECTION OF DATA

Data is collected from web scraping http://agmarknet.gov.in/ with python,and the data is cleaned by removing the unwanted data and changing the type to its appropriate type from the object.

LITERATURE REVIEW

Studied few papers related to the sarima model and vegetable prices,the price farmers receive is rarely fixed and is difficult to predict.So, there exists a need to forecast the tomato prices in the wholesale market hall to evaluate the marketing opportunities in time of all the tomato growers.

Forecasting tomato prices can provide critical and useful information to tomato growers making production and marketing decisions.

The objectives of this paper were to analyse the seasonal price variation of tomato crop and to develop a Seasonal ARIMA (SARIMA) model to forecast the monthly tomato prices at wholesale level in Antalya, a city located in the Mediterranean Region, Turkey, on the basis of reported prices from 2000 to 2010.

DATA DESCRIPTION

The data has 13 columns to represent the cost of tomatoes each day in each market in india.

State - State of the market from where data is collected.

District - District of the market from where data is collected.

Market -Market Name for data.

Variety - It is about the type of tomato(local or hybrid)

Commodity - vegetable

Tons - Number of tons ,market got on a particular day.

Min Price - Minimum price of the day for a ton of tomatoes.

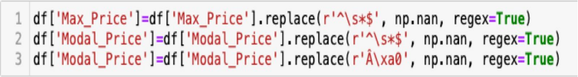
Max Price - Maximum price of the day for a ton of tomatoes.

Modal Price - Average price of the day for a ton of tomatoes.
Arrival Date : Date of the day when data is stored.
Date Month Year

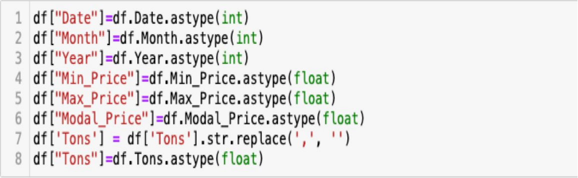
PREPROCESSING THE DATA

- **Removing unwanted data**

As data is web Scraped unwanted data should be removed from the data.

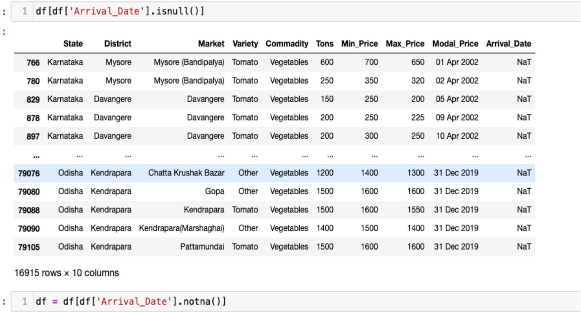


```
df['Max_Price']=df['Max_Price'].replace(r'^\s*$', np.nan, regex=True)
df['Modal_Price']=df['Modal_Price'].replace(r'^\s*$', np.nan, regex=True)
df['Modal_Price']=df['Modal_Price'].replace(r'Â\xa0', np.nan, regex=True)
```



```
df["Date"]=df.Date.astype(int)
df["Month"]=df.Month.astype(int)
df["Year"]=df.Year.astype(int)
df["Min_Price"]=df.Min_Price.astype(float)
df["Max_Price"]=df.Max_Price.astype(float)
df["Modal_Price"]=df.Modal_Price.astype(float)
df['Tons'] = df['Tons'].str.replace(',', '')
df["Tons"]=df.Tons.astype(float)
```

- **Removing NaN values**



```
df[df['Arrival_Date'].isnull()]
```

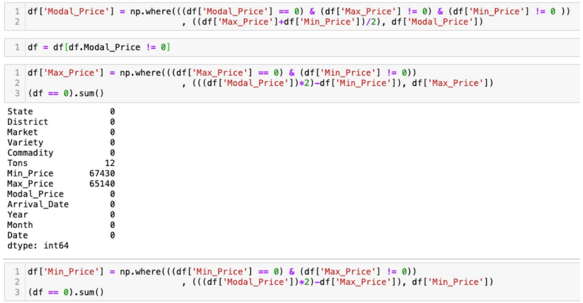
	State	District	Market	Variety	Commadity	Tons	Min_Price	Max_Price	Modal_Price	Arrival_Date
766	Karnataka	Mysore	Mysore (Bandipalya)	Tomato	Vegetables	600	700	650	01 Apr 2002	NaT
780	Karnataka	Mysore	Mysore (Bandipalya)	Tomato	Vegetables	250	350	320	02 Apr 2002	NaT
829	Karnataka	Davangere	Davangere	Tomato	Vegetables	150	250	200	05 Apr 2002	NaT
878	Karnataka	Davangere	Davangere	Tomato	Vegetables	200	250	225	09 Apr 2002	NaT
897	Karnataka	Davangere	Davangere	Tomato	Vegetables	200	300	250	10 Apr 2002	NaT
...	...	...	...	...	...	...	...	...	...	...
79076	Odisha	Kendrapara	Chatta Krushak Bazar	Other	Vegetables	1200	1400	1300	31 Dec 2019	NaT
79080	Odisha	Kendrapara	Gopa	Other	Vegetables	1500	1600	1600	31 Dec 2019	NaT
79088	Odisha	Kendrapara	Kendrapara	Tomato	Vegetables	1500	1600	1550	31 Dec 2019	NaT
79090	Odisha	Kendrapara	Kendrapara(Marshaghai)	Other	Vegetables	1400	1500	1400	31 Dec 2019	NaT
79105	Odisha	Kendrapara	Pattamundai	Tomato	Vegetables	1500	1600	1600	31 Dec 2019	NaT

16915 rows × 10 columns

```
df = df[df['Arrival_Date'].notna()]
```

- **Filling Zeros**

The missing zeros in modal price is filled with the mean of min and max price of the day,and similarly for min price and max price.



```
df['Modal_Price'] = np.where(((df['Modal_Price'] == 0) & (df['Max_Price'] != 0) & (df['Min_Price'] != 0 ))
                             , ((df['Max_Price']+df['Min_Price'])/2), df['Modal_Price'])
```

```
df = df[df.Modal_Price != 0]
```

```
df['Max_Price'] = np.where(((df['Max_Price'] == 0) & (df['Min_Price'] != 0))
                             , (((df['Modal_Price'])*2)-df['Min_Price']), df['Max_Price'])
(df == 0).sum()
```

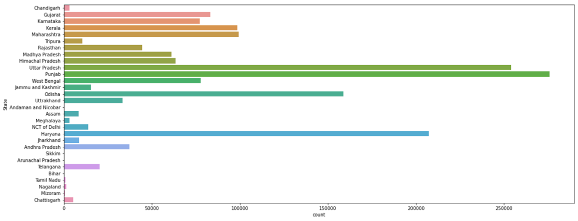
```
State               0
District            0
Market              0
Variety             0
Commadity           0
Tons               12
Min_Price       67430
Max_Price       65140
Modal_Price         0
Arrival_Date        0
Year                0
Month               0
Date                0
dtype: int64
```

```
df['Min_Price'] = np.where(((df['Min_Price'] == 0) & (df['Max_Price'] != 0))
                             , (((df['Modal_Price'])*2)-df['Max_Price']), df['Min_Price'])
(df == 0).sum()
```

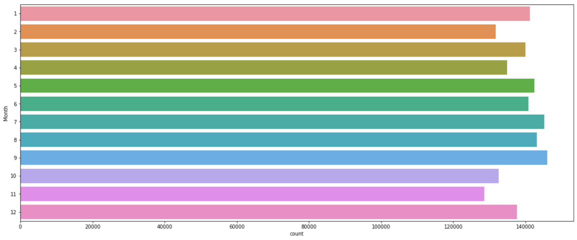
EXPLORATORY DATA ANALYSIS

Considering all the parameters , We have analysed few things from the data.

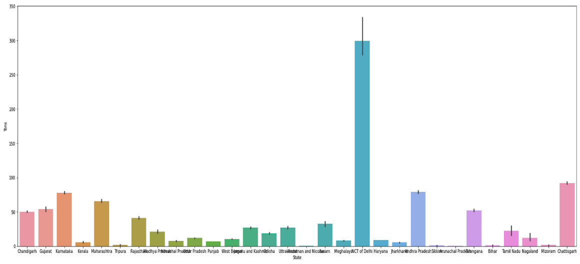
Punjab is the highest producer of tomatoes followed by uttar pradesh and haryana.



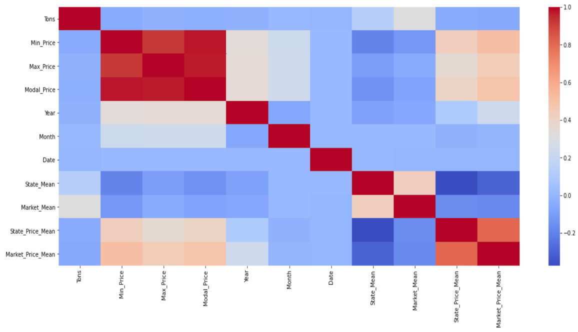
Even though tomatoes are perennial crops ,we can see its grown annually all over the country.



NCT of Delhi markets has got highest tons of tomato



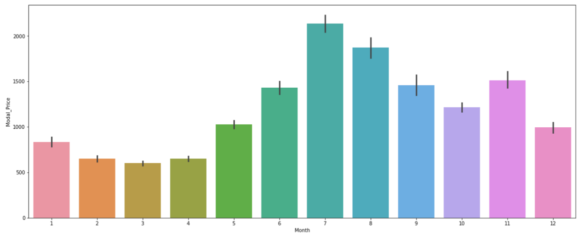
Correlation



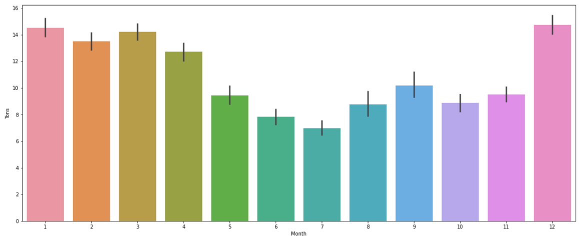
From the correlation graph we can see that the price and tons are correlated.

To Check the correlation properly Ihave plotted the modal price and tons of one market.

Modal_Price



Tons



MODEL BUILDING

To forecast the future prices ,I have decided to build a sarima model as we can see seasonal trend in data, even though tomatoes are annual crops.
To check the data fitting and accuracy of the forecast I tried LSTM to forecast the price for upcoming days.

SARIMA - Seasonal Autoregressive Integrated Moving Average.

LSTM - Long Short Term Memory.

I have considered two arima models

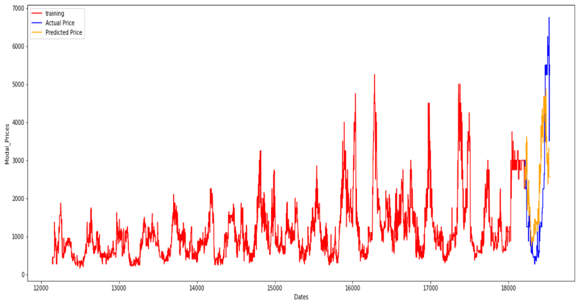
- Considering Model Price as endog with seasonality of 365 days to forecast Model Price.
- Considering Model Price as endog and Tons as exog with seasonality of 365 days to forecast Model Price.

SARIMA

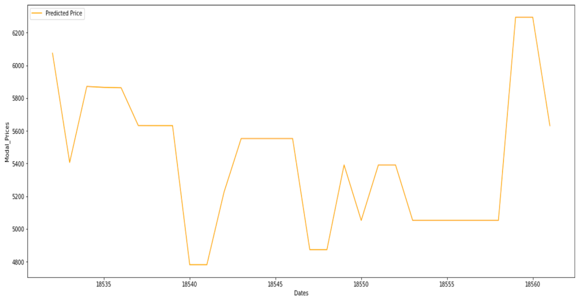
Started with checking if the data is stationary or not , we found data is not stationary and found data is stationary at first differencing,so we concluded d in (p,d,q) to be 1. Then by plotting acf(Autocorrelation function) and pacf (Partial Autocorrelation function) ,we concluded the values of p and q to be 1.

- Only endogenous values

Considering only modal value , splitted data into train data set and test data set,fitted the model with modal price of train data set and used the model to forecast the test dataset and the got rmse of 0.75.
This model can be used to forecast the future price by training the whole data set and forecasting upcoming days.
Testing the model



Forecasting future

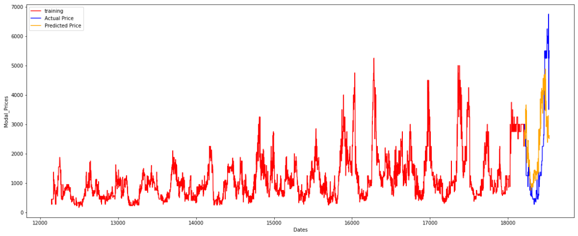


- Considering the exogenous value tons as we have seen from collerogram that modal price and Tons.

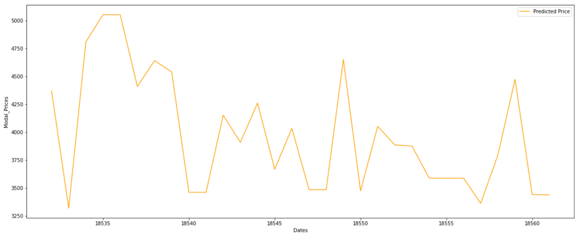
We followed the same procedure by splitting the data into train dataset and test dataset and passed model price of traindata as endog and tons of train data as exog to fit the model ,once the model is trained we used tons from test data to forecast the model price of test data set with rmse value of 0.74

To use this model to forecast future prices ,firstly we build a model with endog for tons to forecast the value of tons for upcoming days and used the same as exog values to forecast modal price for upcoming days.

Testing the model

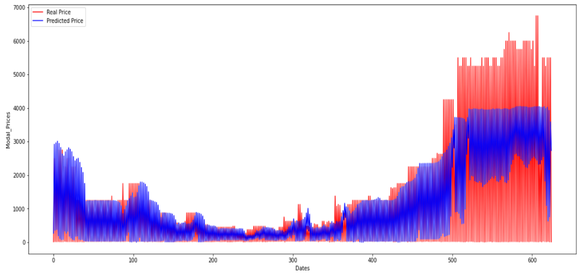


Forecasting Future

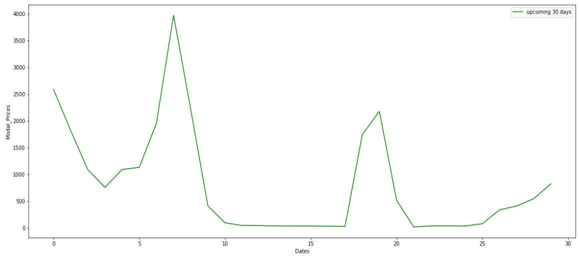


LSTM

We started with splitting the data into train and test data set,transformed the data using a minmax scalar and created a dataset with timesteps of 15 days i.e, to predict the value of the 16th day from analysing the values of 15 days. We fitted the lstm model using keras into 4 layers 3 being lstm layers and the last dense output layer with 20 epochs(more than these caused heavy validation loss) and fitted the model.
We used the fitted model to forecast test data
but the rmse value was too high 586.42 so
we conclude this model can’t be used.
Testing the model



Forecast



RESULTS AND CONCLUSION

From all the above models we can see SARIMA gives accurate values,which is useful for people including farmers, market man and even consumers.
The vegetable price prediction is essential for common people to recognize the price of vegetables in advance. Daily price data of vegetables price data

The decrease of income of tomato growers may bring out the farmer group who is unwilling to continue to produce tomato.Time Series forecasting is really useful when we have to take future decisions or we have to do analysis,we can quickly do that using SARIMA, there are lots of other Models from we can do the time series forecasting but SARIMA is really easy to understand.

FURTHER RESEARCH

As we have seen tons are related to modal price in the same many other factors effect the price such as temperature, rain fall and some other news in the present world(like import export), which can play a major role for the cost.

Also further research can be done by adding more vegetables. More influential factors need to be considered. The work can be further extended using other supervised learning techniques to increase the performance accuracy.

REFERENCES:

https://medium.com/@kfoofw/seasonal-lags-sarima-model-fa671a858729

https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima_model.ARIMA.html

http://www.ijitee.org/wp-content/uploads/papers/v9i2S/B12261292S19.pdf

http://downloads.hindawi.com/journals/mpe/2014/135862.pdf

https://link.springer.com/content/pdf/10.1007%2F978-3-642-18354-6_79.pdf