Ques 1 :

# Python Script to Fetch Data From Jan 1,2020 to Dec 31,2020

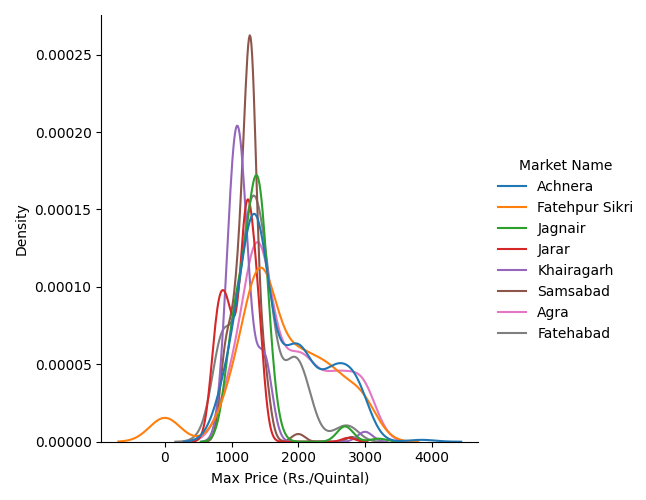
|  |
| --- |
| import pandas as pd  from tqdm import tqdm data = [] month\_day\_dict = {  'Jan':31,  'Feb':29,  'Mar':31,  'Apr':30,  'May':31,  'Jun':30,  'Jul':31,  'Aug':31,  'Sep':30,  'Oct':31,  'Nov':30,  'Mar':31, } for k,v in month\_day\_dict.items():  for i in tqdm(range(1,month\_day\_dict[k]+1)):  url = 'https://agmarknet.gov.in/SearchCmmMkt.aspx?Tx\_Commodity=24&Tx\_State=UP&Tx\_District=1&Tx\_Market=0&DateFrom='+str(i)+'-'+str(k)+'-'+'2020&DateTo='+str(i)+'-'+str(k)+'-'+'2020&Fr\_Date='+str(i)+'-'+str(k)+'-'+'2020&To\_Date='+str(i)+'-'+str(k)+'-'+'2020&Tx\_Trend=0&Tx\_CommodityHead=Potato&Tx\_StateHead=Uttar+Pradesh&Tx\_DistrictHead=Agra&Tx\_MarketHead=--Select--'  # print(url)  try:  values = pd.read\_html(url)  temp = values[0]  # print(temp.values)  # print(temp.head())  for x in temp.values :  data.append(x)  # data.append(temp.values)  # print(data)  except:  print(url)   # Debugging Error # for i in range(len(data)): # if np.shape(data[i])[0]!=10: # print(np.shape(data[i]),data[i],i)  # The data for July 12,2020 is not available and hence causing issue. Manually removing it fixes the error.  data.pop(1201) cols = ['Sl. No','District Name','Market Name','Commodity','Variety','Grade','Min Price (Rs./Quintal)','Max Price (Rs./Quintal)','Modal Price (Rs./Quintal)','Price Date'] df = pd.DataFrame(temp,columns=cols) df.to\_csv('fetched\_data\_v1.csv',index=False) |

# Major Markets in Agra :

|  |  |
| --- | --- |
| Market | Number of Apperances |
| Achnera | 280 |
| Fatehpur Sikri | 265 |
| Agra | 260 |
| Fatehabad | 235 |
| Samsabad | 168 |
| Jagnair | 161 |
| Khairagarh | 157 |
| Jarar | 146 |

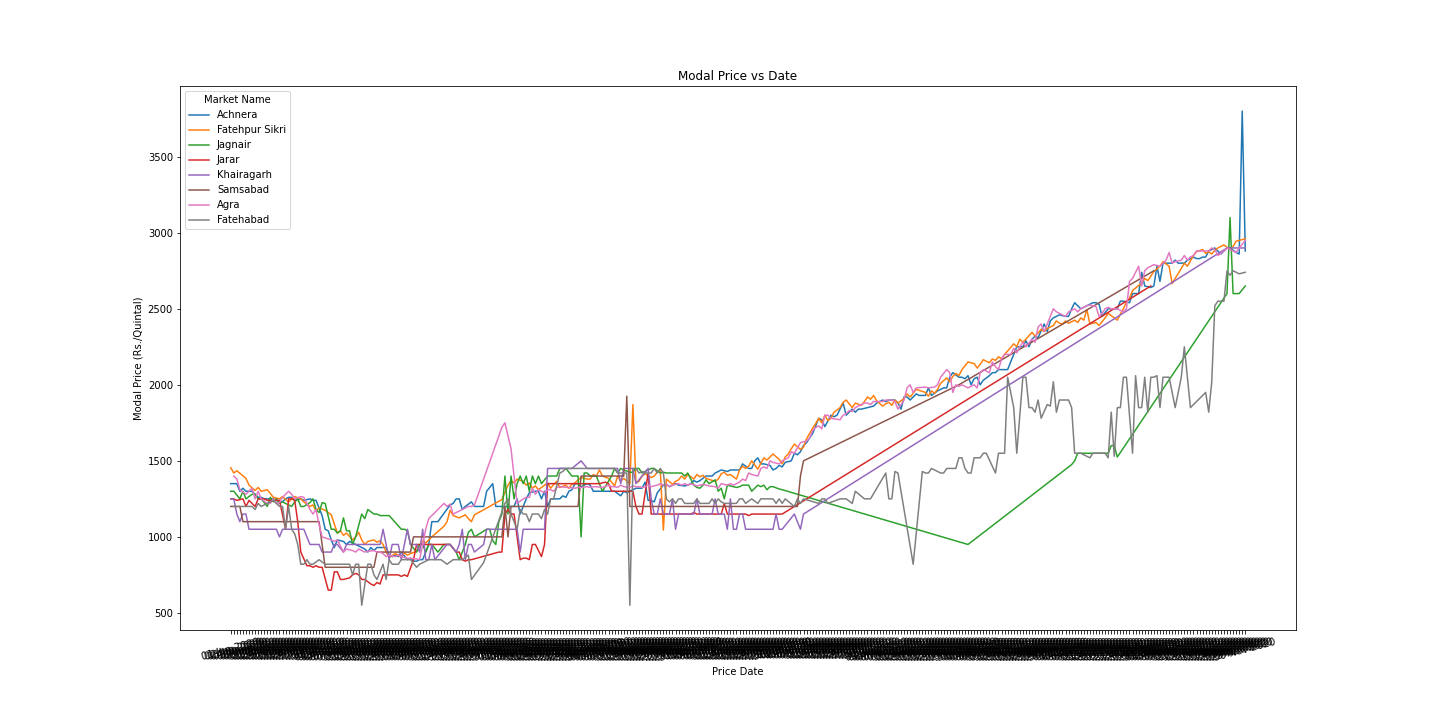
Ques 2 :

## Trends :



* Normal Distribution for almost all the markets.

## Line Plot :



# Patterns :

* The prices of Potato in Agra is usually higher in the winters (November - December)
* Only the top 5 markets have been listed multiple times and have values for most of the year
* There is a gradual increase in the price of Potato from March to November and begins to decline from Jan to March

Ques 3 :

# Machine Learning for Predicting Price of Potato based on Data :

### Data Processing Steps :

* Conversion of Date to extract date features such as Day, Month. Since, we are working with only 1 year of data we did not extract the year but if we had sufficient data we could have used year to predict the trend of data with respect to year
* Since, we have only 8 markets and 4 variety of Potatoes one hot encoding helps the model in distinguishing between them
* Remove unnecessary features such as FAQ, District Name, as they do not contribute to the price and are the same for all the points.
* There is a high correlation between the MIN, MAX price of potato and the best target would be the MODAL Price.
* Removal of MIN and MAX price columns to handle data leakage.

### Features Used :

### Market

### Variety

### Day

### Month

#### Other possible features include :

### Week of Year

### Season

### Machine Learning Framing :

**The following data extracted can be used to predict the estimated price of Potato based on the market, variety and the time of the year.**

* **Regression analysis**
* **Time Series Forecasting**

Since, we extracted data for only one year. Time Series forecasting does not give a desired output and fails to identify the patterns. But if are able to collect data for a longer duration then the following data can be used as a time series prediction machine learning problem.

### Models :

For a quick insight in which models seem to work the best. I ran a pycaret experiment on the data and the following models seem to be working pretty well. AutoML provides a decent insight into the data and which models are working good and which are not which saves a lot of time and effort.

* ExtraTrees Regressor
* Random Forest Regressor
* Decision Tree Regressor

A big advantage of using these models is that unlike Deep Neural Network they are not computationally expensive and can be easily deployed for production.

### Loss Functions used :

* Mean Average Error
* Mean Squared Error
* Root Mean Squared Error
* R2
* Root Mean Squared Log Error

The following are the losses incurred while running AutoML :

### Comments :

Difficulties faced during Data Scraping :

**Approach 1 :** Use the **export to excel** button on the web page to extract the data for the desired dataset. This can be done by simulating a button click with the selected time frame. The following approach did not work since the file generated by the button outputs a corrupted xls file.

**Approach 2 (Worked)** : Recursively iterate over a dictionary of dates and get the values from the table for the particular date and append it.