AGRILINKS ASSIGNMENT

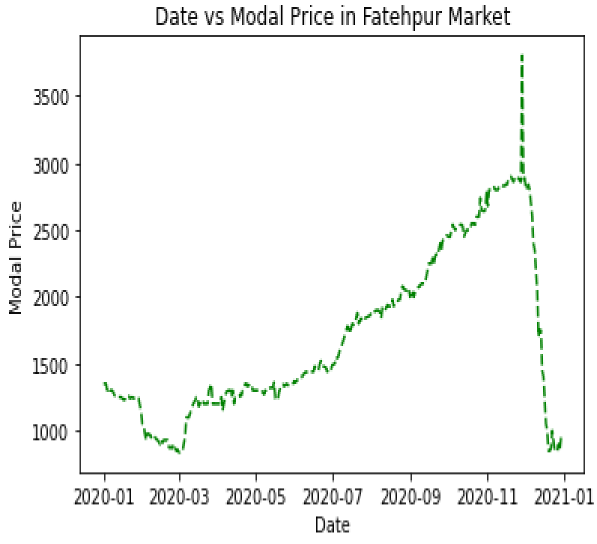
Data Extraction: Collected the data from given url where I used in website filter to get the data belonging to the dates 1 st Jan’2020 to 31 st Dec’2020 in district Agra for crop Potato.
(The script for extracting the above required data is also in the code)

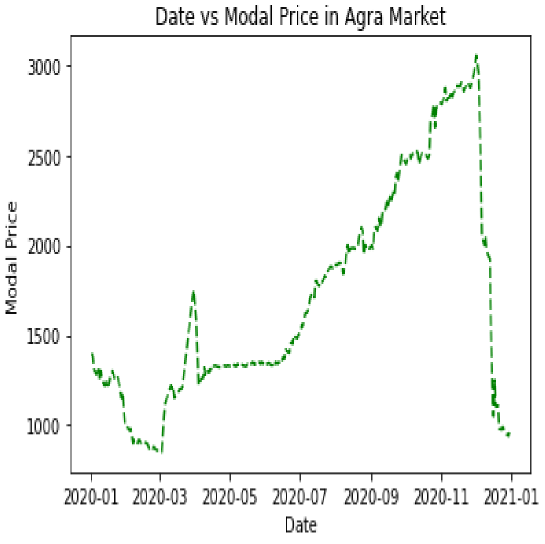
a)
 Please check the ipynb notebook shared
b)

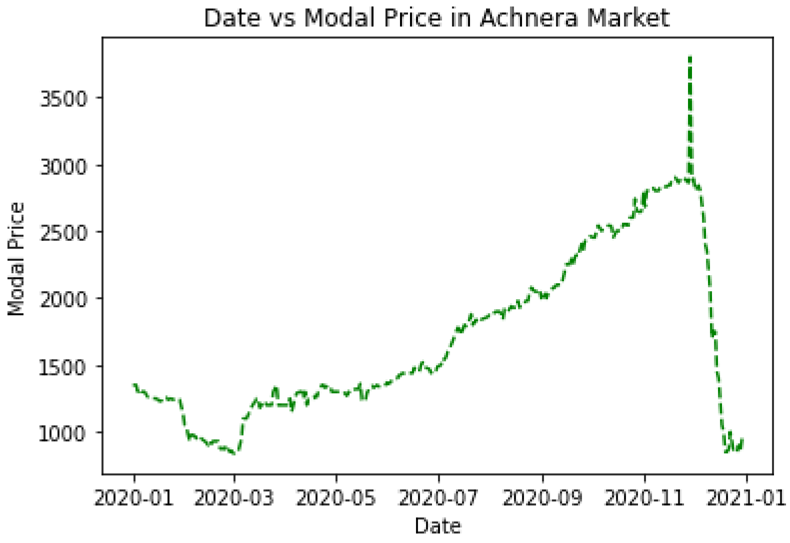
Identifying Major Markets: Group the dataframe based on ‘Market Name’ and find the mean of Modal Price for all dates in 2020.We take top3 Markets with highest mean as Major markets.

```
Market Name         Modal Price Mean
Achnera             1695.212418
Agra                1716.614035
Fatehabad           1385.387597
Fatehpur Sikri      1737.344828
Jagnair             1311.711230
Jarar               1072.671233
Khairagarh          1233.689840
Samsabad            1141.517857
Name: Modal Price (Rs./Quintal), dtype: float64
```

The below plots show the change in Modal price for each day in 2020.







Pattern Identified: All the plots follow a similar pattern we can see the there is slight dip in price from January to March and from that point we can see a steady and steep increase of prices until the start of December and from there there is a steep fall in price across all days of December.

c)

1. Preprocessing/Cleaning techniques:
    a. Handle Missing data either we can Ignore that data or fill in the missing the values based on Population distribution.
    b. Remove outliers as they can trick our loss function
    c. Categorical feature Encoding: The data has categorical features like Market Name, Variety etc. we can either apply One hot encoding or Numerical Encoding as the algorithm can only understand Numerical data.
    d. Data Normalization: We normalize any columns with Numerical data like Max price etc to specific ranges like (-1,1) or (-5,5) because if the values are large algorithm considers the features to have more importance which shouldn’t happen.
    e. Changing Date: Date can also be use as a parameter which can turn out to be important but It is better we use only month as a feature because using year doesn't make sense because a year is not repeated but the price of commodity can vary based on seasons which can be interpreted using months. For example we can transform 01-01-2020 to January which is further encoded as Numerical value.

2. Features used: We can use Market Name, Variety, Month as features to predict Modal price. Further we can plot correlation matrix and remove or linearly combine any highly correlated independent variables. There is no need to use Grade as Feature because it is same for all the entry tuples.
3. Problem Formulation: Given a Market Name, Variety of Commodity and month of selling Predict the Modal price per quintal for the commodity.Here Modal Price is treated as Target variable. This is Regression problem.
4. Algorithm: The preliminary solution would be to use Multi Linear Regression to predict Modal price. Otherwise we can use recent developments li XGBoost for regression tasks.
5. Loss Function: We can use Mean squared error as Loss for our model. However if the data has outliers we can use higher level functions like Huber loss or log-cosh which are used in Xgboost algorithm.