**b) From the plots we observe the following**

* Price pattern in Samsabad market had price fluctuations between 700 – 1500 and is on an increasing trend after July 2020.
* In the Achnera market the prices were on increasing trend until 30 November 2020. With a sudden spike on Nov 30, later the prices are in decreasing pattern until December 30.
* In Jagnair market prices ranged between 900 – 1500 until August 31, 2020. And the prices spiked to 3100 until Nov 25, 2020 and got a sudden decrease later.
* In Khairagarh market prices seem to be fluctuating a little bit until June 7, 2020. Then there was an increasing trend pattern until Nov 30, 2020 and got a sudden drop later.
* The price pattern in Fatehpur Sikri is bit interesting and there is too much variations between maximum, minimum prices and model prices. Model price seem to be constant. This may be indicating the outliers in Max and Min price.
* In Jarar market we see the increasing trend from July 1 to end of the year with the highest price of 2800 Rs.
* We observe increasing trend in Agra market until December 2 and a sudden drop in prices further
* Fatehabad market shows a little fluctuation in the prices with Max price of 2800 and Min price of 550.
* The common point we were able to observe that the prices drop by the end of the year especially in the month of December of sees Maximum by the end of November also the prices tend to increase after the month of July across all the markets.

**c)**

**What are the data pre-processing / cleaning techniques you would apply?**

* Removing the columns such as Sl no., District Name, Commodity, Grade
* Renaming columns
* Converting Price date to date-time format
* Sorting the data in ascending order by Price Date and resetting index
* Label encoding the string columns
* Extracting the features from columns
* removing the price date and making the data ready for modelling
* Scaling the dataset

**What are the features you would use to create the model?**

 Minimum price

 Maximum price

 market name

 Month

 is\_holiday

 Day\_of\_week

 Variety (remove potato and other)

 Weekend or not

 Quarter

 Season

 Week\_of\_year

 Difference price between Min and Max price

 Difference price between Modal and Max price

 Difference price between Modal and Min price

 Modal price (Target variable)

**How would you frame this problem as a machine learning problem? What would be the target variable?**

This problem can be solved in 2 ways

* Classical Machine Learning
* Time series

Here Modal price can be chosen as the **target variable.**

In Machine learning we can create the model by using the above features and based on the performance on cross validation we can add or remove the features. Here since the data is time indexed, we need to split the data for the training and validation sequentially. I would also add some extra features related to currency value and weather for the market location, because weather can also be one of the important factors influencing the price.

In Time series method Modal price with date index can be used to create model, further we can create models for forecasting the minimum and maximum prices for different markets using different models such as ARIMA, LSTM and Prophet.

**Which algorithm would you use for price prediction?**

I would go with random forest and gradient boosting because, in our data most of the features are categorical, so I think tree-based model would perform better. Also I would look into the ensemble of ML and time series predictions.

**What would be the loss function you would use?**

I would use Root mean square error as my loss function