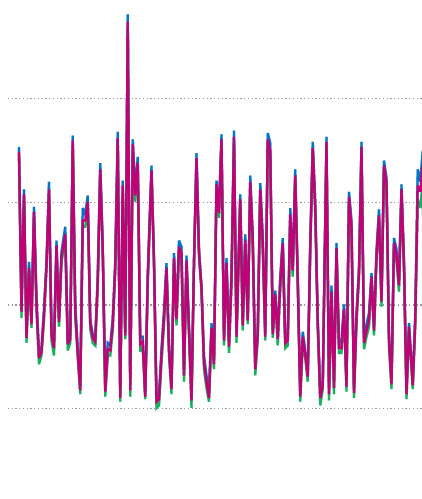
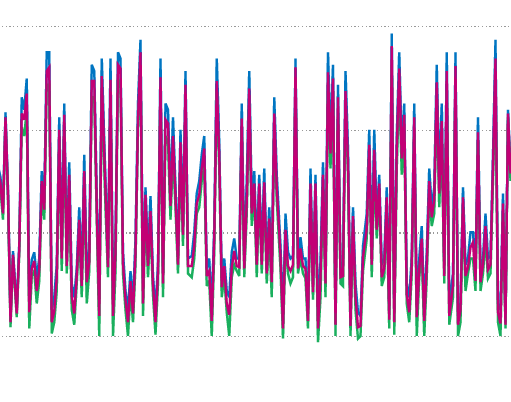
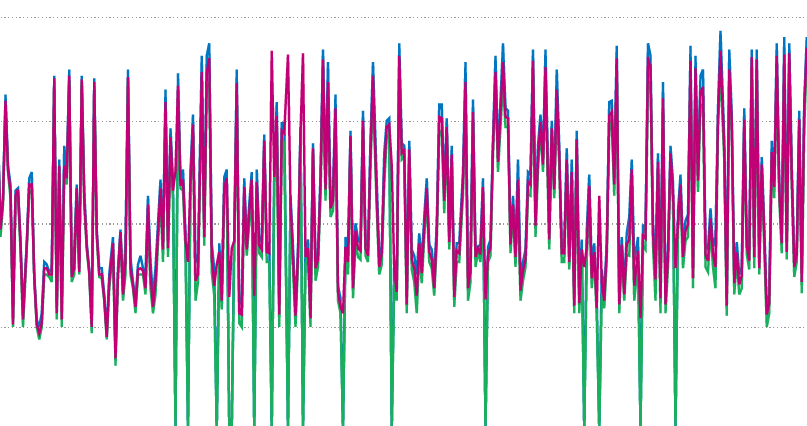
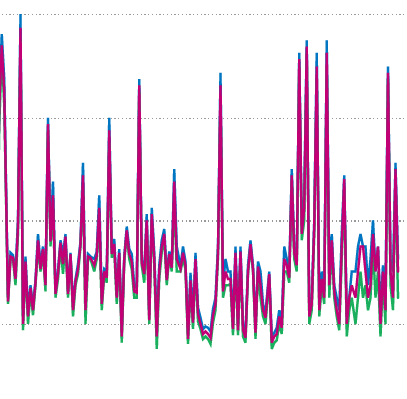
**Part B**

***1.Identify major markets for the district “Agra” and plot price patterns for each of them. What patterns do you identify?***

**ACHENERA**   **AGRA**

**Fathehbad and Fatehpur Sikri**  **Jagniar and Others**

* I have broken the market plots into four parts based on their trends in four parts, based on their valence.
* We can see that in the first **ACHENERA** the Modal values are going a bit higher, hence we can see that the variance is high.
* In the next plot of AGRA, we can visualize clear and similar variance in each of the points so, this is another kind of market we can see.
* In the last one Fathebad and Fatehpur Sikri, the min price is very low, in all the plots this is the least and this another kind of market trend.
* In the other markets in some areas, the modal prize is low means the max prize and min prize both is low. So this another kind of market trend.

**Part C**

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

* First, we should make individual datasets different datasets based on their markets, it will be easier for us to analyze the markets and also have a good understanding of the dataset.
* Then we can remove the SL\_Name, District Name, Commodity, grade and market name, the reason is SL\_name is an index value we shouldn’t put that in the dataset, the District Name, Commodity, grade and market name will be the same for each of the individual datasets.
* We should put the min prize, max prize because that can be our target value(one of them) and the model prize shouldn’t be required as the prizes are already given.
* We can change the date to date time format and use it for different kind of models like time series forecasting or other regression techniques. We also have to sort that based on the date format of the calendar.
* We can also remove the prize outliers as they can affect the model, but if we using the time series forecasting for that we need to put all the data because that may affect the seasonality.
* We also have to check the missing values in the min prize and the max prize, if there are any missing values we should fill or replace them with the mean or median of the prize(min/max) of that particular month, we shouldn’t go for the mean of the whole year as the seasonality may vary month to month, so it should be better to put the median or mean of a month data.

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

* For a regression task, we need more features, after data cleaning we are remaining with Variety, max & min prize and the date-time column.
* **Difference between max and min:** For each of the market dataset, we can have another data which will be the difference between the max and mean price.
* Month: For a regression task the data should be classified and the best correlation we can find out with the months for the date-time we can find out the months’ column and which can help us to predict and get a good score.
* **Week number:** we can also add the week number and there will 4 classes only the model will get confidence to predict from the datasets, and it can give us better results too.
* Other than this we can also find the feature importance and from that, we can also see how is the status of the feature importance score.

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

* We particularly frame this problem statement as a regression task and from that, we can utilize the prediction for normal regression the data should be clean and there should be any outliers too, but for the case of forecasting, it will be different.
* Target: We can take the Min Prize or Max precise any of the target value and from that, we will understand the prediction of the future.

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

This particular problem can be solved in two ways.

1. Time series forecasting: This particular dataset is perfect for time series forecasting, basic Arima or FB Prophet we can use to forecast, for that we only need the date and that should be the index we can understand the seasonality and each day data rows should be present. Based on the seasonality we can understand the data and make the model and from that, we can make a good prediction.
2. Regressional analysis: In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

As discussed earlier we have to use those features to predict and make a good model. We can use basic Regessor algorithms and based on the best results we can stack them using a stacking regressor.

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

We for time series forecasting we can use future data to test the model and find the score and for the basic regressors, we can split that data and test it. We can use the two scoring methods.

* R2 score: In statistics, the coefficient of determination, denoted R² or r² and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable.
* RMSE: The root-mean-square deviation or root-mean-square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

***6.Any other comments you want to add?***

* This particular problem will be the best time series forecasting problem statement because the variety of the potato can vary in another year.
* If there are some other features added then we can also take this problem in a basic statistical regression task. Let us discuss them.
  + The much potato of each variety has grown in each month and how much is coming to the market based on that we can easily do the analysis, as the more potato come means more supply will help the people to buy as the prize will go down.
  + The population around that area of the market is also important because that affects the prize too, as the more population the prize will go high.
  + Another will be a public review, the board which is checking the quality of the commodity they give reviews based on them the price of the commodity may vary.