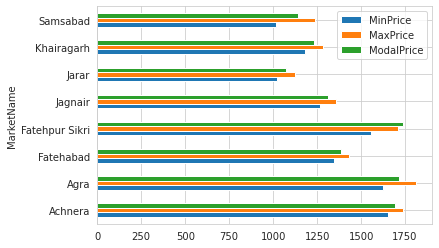
**GRAMODAY INTERNSHIP ASSIGNMENT**

**Task 2 : Identify major markets for the district “Agra” and plot price patterns for each of them. What patterns do you identify?**

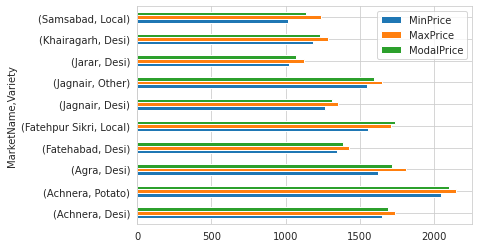
1. There are 8 Markets ( 'Achnera', 'Agra', 'Fatehabad', 'Fatehpur Sikri', 'Jagnair', 'Jarar', 'Khairagarh', 'Samsabad' ) in Agra District where Potato crop is sold.
2. Plot of Mean Price(for all 3 Price Categories) vs Market  
   

Here,

1. Average of all three prices was high for Achnera, Agra and Fatehpur Sikri Markets in comparison to other Markets in 2020.

2. Price Averages were low for Jarar, Khairgarh and Samsabad in 2020.

1. Plot of Mean Price(for all 3 Price Categories) vs Market + Variety

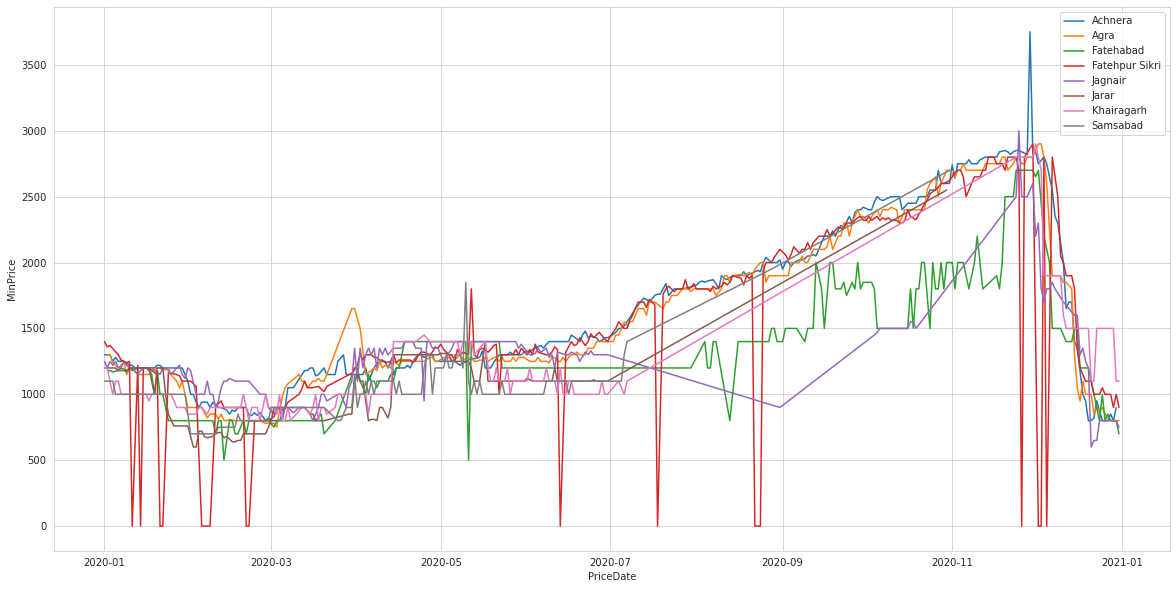
  
Here,

1. Looks like not all varieties are sold at all Markets.

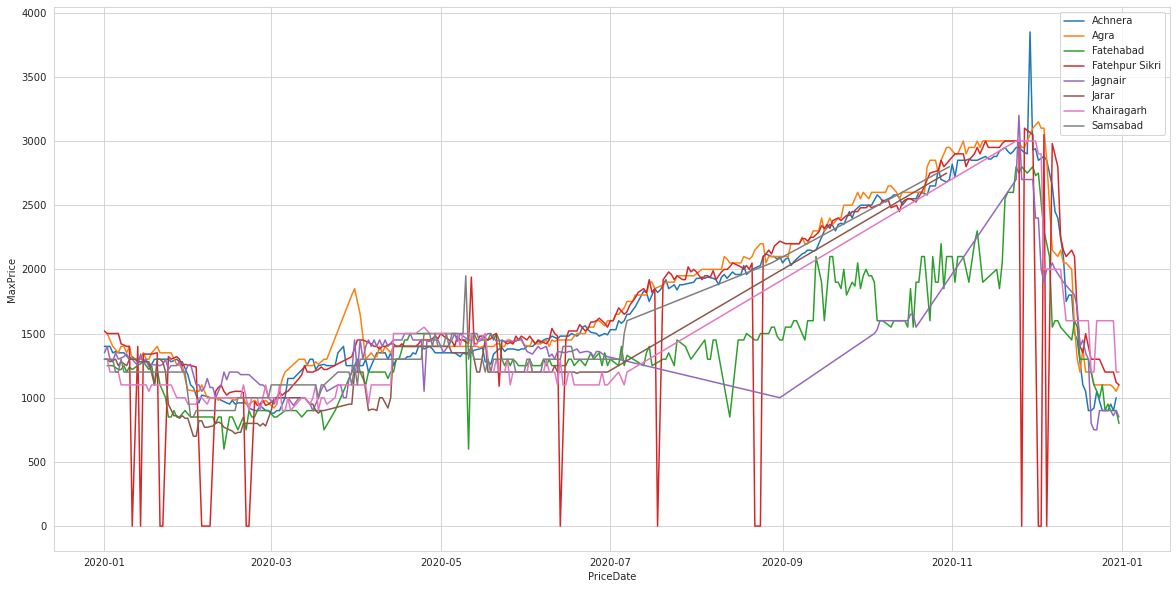
2. Mean Prices are highest for ‘Potato’ Category at Achnera Market.

3. Mean Prices are highest for ‘Desi’ Category at Jarar Market.

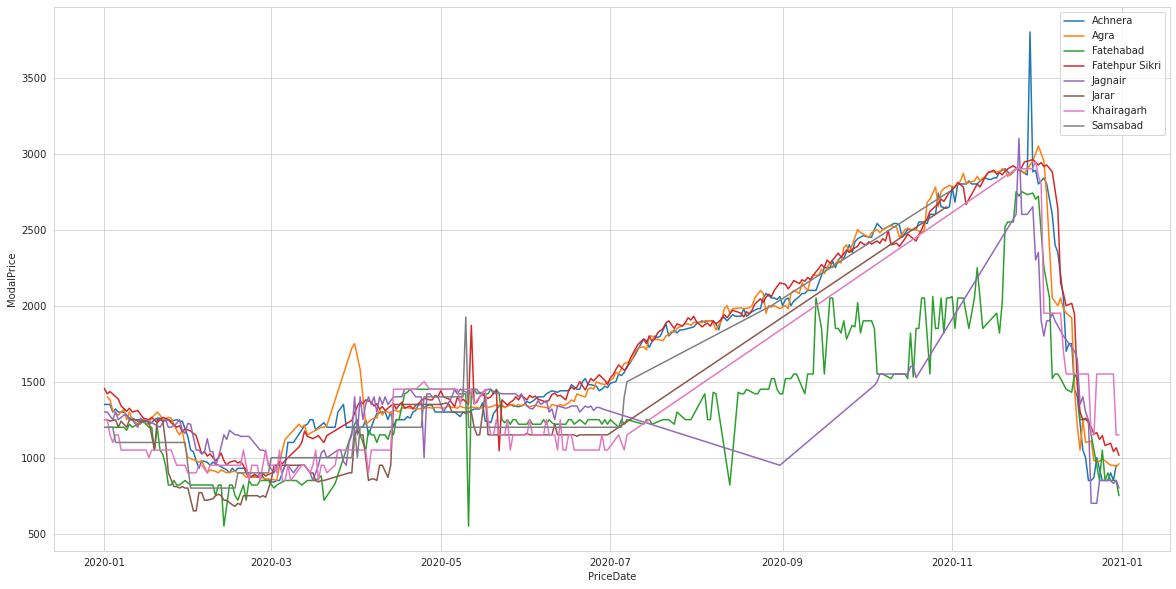
1. Market-wise Prices vs Price Date

1. MinPrice vs Date  


2. MaxPrice vs Date



3. ModelPrice vs Date



In the above graphs,

1. In case of all prices, we see an increasing trend startinng from Sept, 2020.
2. And a steep decreasing trend is observed starting from Dec, 2020.
3. Prices are 0 for some days in Fatehpur Sikri Market. Might be due to Holidays or these represent NaNs.
4. There is high fluctuation in prices at Fatehabad Market.

**TASK C - Comment on how you can leverage machine learning to predict prices for a given market in Agra for the crop “Potato”.**

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

Ans : Preprocessing Steps:

1. First we need to check and set the data types of the feature columns. Scraping data from websites usually returns string values. Convert the values to the appropriate data type.
2. Check for missing values. If there are any, use median, mode or fit a linear regression model to predict the missing value.

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

1. As we have a time series problem, I’d focus more on creating lag features. In lag features we shift the feature values down to add additional info about past data in the current row.
2. Rolling Window Average Features with different batch sizes will also help us capture the trend of prices in a short window.
3. I’ll also use the target ‘Price’ column as a feature by adding lag.
4. Also, we can add previous month aggregates as features in the current month rows.

iii. How would you frame this problem as a machine learning problem? What

would be the target variable?

1. I’d frame this as a Time Series problem. And will try two Cross-Validation Methods: KFold and TimeSeriesSplit.
2. We can create 2 models with targets MinPrice and MaxPrice as the features will be same for both.

iv. Which algorithm would you use for price prediction?

1. If we are predicting price for a single market, we won’t have much data to train the model. In this situation, I’d avoid decision tree based model as they are more likely to overfit.
2. I think linear models with regularization will be a good choice. I’d go for Ridge Regression.

v. What would be the loss function you would use?

1. We can use regression losses like Mean Squared Error, Root Mean Squared Error, Mean Absolute Error. I’d choose Mean Absolute Error as it’ll be more intuitive in this case.

vi. Any other comments you want to add?

1. Nothing as such. Thanks for this interesting assignment. I really enjoyed working on it.